REPURPOSING DIGITAL METHODS

The research affordances of platforms and engines

Esther Weltevrede

Repurposing digital methods

The research affordances of platforms and engines

Academisch proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op 28 januari 2016, te 10.00 uur

door

Esther Josephina Theresia Weltevrede
geboren te Geldermalsen

Promotiecommissie

Promotor:	Prof. dr. R.A. Rogers	Universiteit van Amsterdam
Overige leden:	Prof. dr. L.W.M. Bod	Universiteit van Amsterdam
	Prof. dr. R. Boast	Universiteit van Amsterdam
	Prof. dr. A. Bruns	Queensland University of Technology
	Prof. dr. M.J.P. Deuze	Universiteit van Amsterdam
	Prof. dr. H. Kennedy	University of Sheffield
	Prof. dr. J.J. Noordegraaf	Universiteit van Amsterdam

Faculteit: Geesteswetenschappen

The research of this doctoral dissertation received financial support from the Amsterdam School of Cultural Analysis (ASCA PhD fellowship) and two European grants for the Contropedia project (EU FP7 EINS grant #288021 and in part EU FP7 EMAPS grant #288964).

Table of Contents

Acknowledgements

This dissertation would not have been possible without the support of my family, my partner and son, my friends and colleagues and the many people I have had the pleasure to meet along the way, some of whom I'd like to mention here.

First, I want to thank my supervisor Richard Rogers for his support, inspiration and for being the best supervisor I could wish for. Richard has been an encouraging mentor, always pushing my ideas, and an inspiring advisor, whose thought-provoking insights made the experience of writing this thesis a joyful experience. I am extremely grateful for being part of the creation and development of the Digital Methods Initiative since the summer of 2007. I consider this research group my intellectual home. Above all, DMI is a productive, fun, challenging and tightly-knitted group of people, which I consider among my closest friends. Thank you: Richard Rogers, Sabine Niederer, Erik Borra, Bernhard Rieder, Anne Helmond, Michael Stevenson, Carolin Gerlitz, Lonneke van der Velden, Marc Tuters, Natalia Sanchez Querubin, Simeona Petkova, Nadia Dresscher-Lambertus, Liliana Bounegru, Jonathan Gray, Saskia Kok, Emile den Tex, Koen Martens, Anat Ben-David, Marieke van Dijk, Auke Touwlager, Bram Nijhof, Laura van der Vlies, Rosa Menkman and Marijn de Vries Hoogerwerff. My work benefited greatly from the encounters, collaborations and the discussions I had with all of them.

I especially want to thank Sabine Niederer for her friendship, her keen eye for good ideas and positive and stimulating advice. Thank you dear Anne Helmond, whom I first met when we were students at the New Media program, for being greatly inspiring and medium-specific. I am also grateful to Carolin Gerlitz, whom I first met during the 2009 Summer School, for her friendship, inspiring collaborations and useful comments on chapters of this book. I also want to thank Bernhard Rieder for his humor and stimulating conversations and my friends from early DMI, Michael Stevenson, Rosa Menkman and Marijn de Vries-Hoogerwerff for their humor, friendship

and expertise. I also would like to thank Noortje Marres for being such an inspiration and friend. Especially our close collaboration on the Scraping the Social article has been key to developing the larger argument of this book. These many collaborations and friendships have been instrumental in developing thoughts and methods in this dissertation.

I also would like to thank all the participants of the DMI Summer and Winter Schools that I had the pleasure to meet throughout the years for the interesting discussions, collaborations and feedback, as well as the collaboration on the various empirical projects throughout the dissertation. I especially want to thank David Moats and Johannes Passmann for providing me with helpful comments on chapters during Winter Schools. Thanks to Fieke Jansen, Sami Ben Gharbia, Demet Dagdelen, Martin Feuz, Marije Rooze, Thomas Poell, Taina Bucher, Vera Bekema, Andrea Fiore, Simon Marschall, Elena Tiis, Cameran Ashraf, Bronwen Robertson, Leva Zand and Niaz Zarrinbakhsh for their inspiring collaborations on empirical research projects. I am grateful to them all.

A special thanks to the Contropedia consortium for the many productive workshops across the beautiful cities of Europe. Thank you: Tommaso Venturini, Paul Girard, Mathieu Jacomy, Andreas Kaltenbrunner, David Laniado, Paolo Ciuccarelli, Michele Mauri and Giovanni Magni. Thanks to Fieke Jansen and Nishant Shah for their stimulating collaborations, events and co-publications at Hivos and beyond. I am thankful for the hospitality shown by the Iran Media Program at the Annenberg School of Communication, University of Pennsylvania and the inspiring collaboration with Ebby Sharifi. I also want to express my thanks to my Italian friends at the Paolo Alto Winter School for our discussions, collaborations and helping me develop my thoughts about the contribution of digital methods to STS-informed social research. A special thanks to: Massimiano Bucchi, Costanzo Ranci, Trevor Pinch, Matteo Tarantino, Andrea Lorenzet, Simone Tosoni, Marta Severo and Donato Ricci.

Thanks to the Media Studies staff at the University of Amsterdam, and the New Media & Digital Culture staff in particular, for being such pleasant colleagues. I specifically want to thank José van Dijck for her advice on my PhD proposal. Special thanks to Geert Lovink and the people at the Institute of Network Cultures for offering me the opportunity in a very early stage to present my work at one of the new media conferences. Thanks to the Amsterdam School for Cultural Analysis, the research school that saw potential in me and my proposal for funding my PhD, and for their support throughout the process. Thank you Eloe Kingma, Esther Peeren, Christoph Lindner and Wanda Strauven.

Thanks to Mathieu Jacomy for his tools and feedback, Jan–Willem Hiddink and Robert–Reinder Nederhoed for providing a database dump from Loglijst. I would like to thank Anneke Agema for their editorial advice on an earlier version of Chapter 4. The chapters have been presented at various conferences and I thank the conference participants for their insightful questions and suggestions. In addition, I wish to thank the anonymous reviewers and editors from *Journal of Cultural Economy*, *Theory Culture & Society*, *First Monday*, *La Médiatisation de l'Évaluation / Evaluation in the Media* and *A Companion to New Media Dynamics* for their feedback and suggestions.

I would like to give my sincere thanks to Marguerite Lely for her editorial advice and useful comments, which were extremely helpful in improving the quality and clarity of the language. A special thank you to Michele Mauri for the beautiful design of this dissertation.

Finally, I would like to thank my family. A special thanks to my parents Geva and Pieter Weltevrede, my brothers Joel and Nona, my in-laws Marguerite Lely and Dirk Wouters and my sister and brother-in-law Vere Borra and Stef Robijn for their unconditional support and encouragement. A warm thank you to my dearest Milo for being a joy to be around everyday and for his relentless pleasure in exploring and learning new things. Most of all, I want to thank my partner Erik Borra. Thank you for always challenging and inspiring me, for always building tools, for our productive collaborations and for being the best partner in life I could wish for.

Acknowledgements co-authored articles

Marres, Noortje, and Esther Weltevrede. 2013. “Scraping the Social? Issues in Live Social Research.” Journal of Cultural Economy 6 (3): 313–35.

Chapter 2 is based on a two-year collaboration (2011-2013) with Noortje Marres, Goldsmiths, University of London. The chapter was presented in different versions on several occasions, including the Evaluation in the Media colloquium (2012) in Paris and the Digital Methods Winter School mini-conference (2013) in Amsterdam. Another version of the article also appeared as Marres, Noortje and Esther Weltevrede. 2015. “Scraping the Social? Issues in Real-time Social Research,” in Julie Bouchard, Étienne Candel, Hélène Cardy and Gustavo Gomez-Mejia (eds.) *La Médiatisation de l’Évaluation / Evaluation in the Media.* Peter Lang.

After teaching workshops and courses together on scraping and crawling, and discovering our shared interest in questions relating to repurposing digital media for digital research, Noortje Marres invited me to collaborate with her to conceptualize some of the empirical work we had pursued. The empirical work for this study was carried out working from a distance, Noortje from London and I from Amsterdam. We used Skype, email and Dropbox to share and discuss work. The text was written collaboratively by exchanging Word documents and is equally attributed to both authors. This chapter was rewritten to add calling APIs as a distinct collection technique and to focus on the device-driven research perspective.

Weltevrede, Esther, Anne Helmond and Carolin Gerlitz. 2014. "The Politics of Real-time: A Device Perspective on Social Media Platforms and Search Engines." Theory, Culture & Society 31(6): 125–150.

Chapter 3 is based on a four-year collaboration (2010-2014) with Anne Helmond, University of Amsterdam and Carolin Gerlitz, University of Amsterdam. This paper was presented in different versions at several conferences, including the Digital Methods Winter School mini-conference (2013) and the Social Media and the Transformation of Public Space conference (2014), both in Amsterdam.

The paper resulted from our work at the Digital Methods Summer School 2010 at the University of Amsterdam, where I led a project week on web temporalities. This led to the 'Pace Online' project, aka 'One Day on the Internet Is Enough', with Erik Borra, Taina Bucher, Carolin Gerlitz and Anne Helmond and I would like to thank my colleagues for thinking through this project, which eventually resulted in "The Politics of Realtime" article co-authored with Anne and Carolin. The article was written collaboratively in Google Docs and is equally attributed to the three authors. Adjustments were made to focus on the contribution of the empirical work as digital methods for software studies, which I refer to as medium research.

Weltevrede, Esther and Anne Helmond. 2012. "Where Do Bloggers Blog? Platform Transitions within the Historical Dutch Blogosphere." First Monday 17(2).

Chapter 4 is based on a two-year collaboration (2011-2013) with Anne Helmond, University of Amsterdam, and the empirical work was done in collaboration with Erik Borra, University of Amsterdam (2011). The chapter was presented in different versions on several occasions, including the Out of the Box: Building and Using Web Archives conference (2011) organized by the International Internet Preservation Consortium at the National Library of the Netherlands in The Hague and the MiT7 Unstable Platforms conference (2011) at the Massachusetts Institute for Technology in Cambridge, MA.

The empirical study was initiated in Barcelona, where I was living at the time to work full-time on the dissertation. The empirical work was commenced by me working from Barcelona with Anne working from Amsterdam, followed by sessions carried out together in Barcelona (Spring 2011) and in Amsterdam (2012-2013). Erik contributed to the project by building custom tools. The text was written collaboratively in Google

Docs and is equally attributed to both authors. This chapter was rewritten focusing on the volatility of web technology and the device-driven perspective in digital research.

Weltevrede, Esther and Erik Borra. 2015. "Controversy in the Back-end of Neutral Point of View: The Research Affordances of Wikipedia for Studying Societal Issues." Unpublished ms.

Chapter 6 is based on an ongoing collaboration (2012-), initially started with Erik Borra, University of Amsterdam, and since 2013 it is an ongoing European collaboration with MédiaLab SciencesPo (Tommaso Venturini, Paul Girard, Mathieu Jacomy), Barcelona Media (Andreas Kaltenbrunner, David Laniado), Density Design (Paolo Ciuccarelli, Michele Mauri, Giovanni Magni) and the Digital Methods Initiative (DMI) (Richard Rogers). The Contropedia Consortium received a Network of Excellence in InterNet Science (EINS) grant in 2013. This paper was presented in different versions at several conferences, including the Public Communication of Science and Technology conference (2012) in Florence, the Digital Methods Winter School mini-conference (2013) in Amsterdam and the KNAW mini-symposium in a panel with Wikipedia's co-founder Jimmy Wales (2015) in Amsterdam. Different versions were subsequently also published as conference proceedings: Borra, Erik, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers and Tommaso Venturini. 2015. "Societal Controversies in Wikipedia Articles." CHI'15 - Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems: 193-196; and as Borra, Erik, David Laniado, Esther Weltevrede, Michele Mauri, Giovanni Magni, Tommaso Venturini, Paolo Ciuccarelli, Richard Rogers and Andreas Kaltenbrunner. 2015. "A Platform for Visually Exploring the Development of Wikipedia Articles." ICWSM '15 - Proceedings of the 9th International AAAI Conference on Web and Social Media.

The paper resulted from a project idea that had been developing for a while and was inspired by several Wikipedia- related projects from DMI. I would like to thank both my colleagues at DMI and at the Contropedia consortium for thinking through this project. The Contropedia project, which initially was called Controversy Research with Wikipedia, resulted in the unpublished manuscript co-authored with Erik, which is currently submitted to a high-quality journal. The paper was written collaboratively in Google Docs and is equally attributed to both authors. Adjustments were made to focus on the device cultures perspective.

Rogers, Richard, Esther Weltevrede, Sabine Niederer, and Erik Borra. 2013. "National Web Studies: The Case of Iran Online." In Hartley, John, Axel Bruns and Jean Burgess (eds.) A Companion to New Media Dynamics, 142–66. Oxford, UK: Wiley-Blackwell.

Chapter 7 is based on a two-year collaboration (2011-2013) with Richard Rogers, University of Amsterdam, Sabine Niederer, University of Applied Sciences Amsterdam and Erik Borra, University of Amsterdam. Another version of the article also appeared as Rogers, Richard, Esther Weltevrede, Sabine Niederer, and Erik Borra. 2012. National Web Studies: Mapping Iran Online. Philadelphia: Iran Media Program, Annenberg School for Communication, University of Pennsylvania and as an adapted chapter in Digital Methods (2013b) by Richard Rogers.

The paper resulted from an invitation by the Iran Media Program at the Annenberg School of Communication, University of Pennsylvania to Richard Rogers and Sabine Niederer to map the Iranian web using a digital methods approach. Richard and Sabine subsequently invited Erik Borra and myself to join the project. The empirical project was initiated in collaboration with Ebby Sharifi (2011) and continued during the DMI Summer School (2011) at the University of Amsterdam in a project led by me in collaboration with Cameran Ashraf, Bronwen Robertson, Leva Zand and Niaz Zarrinbakhsh. The empirical research was thereafter finalized by Richard Rogers, Sabine Niederer, Erik Borra and myself. I would like to thank all my colleagues for thinking through this project, the Iran Media Program for extending the invitation and the Citizen Lab for their valuable feedback on the research. The text was written collaboratively in Google Docs. We collaboratively produced the paper, but not all parts are attributed equally. Richard Rogers focused on positioning the work, I wrote the main parts of the case study and the findings, Sabine focused on the Fiery Language part of the case study (not included in the chapter), and Erik focused on the technical specificities. This article was rewritten as Chapter 7 focusing on the comparative device cultures aspect of the study as a contribution to censorship research.

List of figures

List of tables

Referenced tools

Censorship Explorer. 2012 Developed by Erik Borra, Emile Den Tex and Richard Rogers for the Digital Methods Initiative. http://tools.digitalmethods.net/beta/proxies/ [Accessed August 30, 2015].

Contropedia Demo. 2015. Developed by Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini for the Contropedia consortium. http://contropedia.net/demo [Accessed August 30, 2015].

DMI-TCAT. 2014. Developed by Erik Borra, Bernhard Rieder and Emile Den Tex for the Digital Methods Initiative. https://github.com/digitalmethodsinitiative/dmi-tcat [Accessed August 30, 2015].

G-Atlas. 2011. Developed by Mathieu Jacomy for TIC Migrations. http://www.e-diasporas.fr/ [Accessed August 30, 2015].

Gephi. 2013. Developed by Mathieu Bastian, Sebastien Heymann and Mathieu Jacomy for Médialab Sciences-Po. https://gephi.github.io/ [Accessed August 30, 2015].

Google Scraper. 2007. Developed by Erik Borra, Koen Martens, Emile Den Tex, Richard Rogers, Sabine Niederer and Esther Weltevrede for the Digital Methods Initiative. https://tools.digitalmethods.net/beta/scrapeGoogle/ [Accessed August 30, 2015].

Internet Archive Wayback Machine Link Ripper. 2009. Developed by Erik Borra, Esther zvWeltevrede, Anne Helmond, Michael Stevenson, Marijn De Vries Hoogerwerff and Richard Rogers for the Digital Methods Initiative. https://tools.

digitalmethods.net/beta/internetArchiveWaybackMachineLinkRipper [Accessed August 30, 2015].

Issue Crawler. 2002. Developed by Richard Rogers, Noortje Marres, David Heath, Suzi Wells, Marieke van Dijk, Auke Touwslager, Erik Borra, Koen Martens and Andrei Mogoutov for Govcom.org. https://issuecrawler.net [Accessed August 30, 2015].

OpenRefine. 2013. Developed by David Huynh for Metaweb Technologies, Inc. http://openrefine.org [Accessed August 30, 2015].

Source Code Search. 2012. Developed by Erik Borra, Esther Weltevrede and Anne Helmond for the Digital Methods Initiative. http://tools.digitalmethods.net/beta/sourceCodeSearch [Accessed August 30, 2015].

TLD Counts. 2013. Developed by Erik Borra and Emile Den Tex for the Digital Methods Initiative. https://tools.digitalmethods.net/beta/tldCounts/ [Accessed August 30, 2015].

Tracker Tracker. 2011. Developed by Koen Martens, Emile Den Tex, Erik Borra, Esther Weltevrede, Anne Helmond and Carolin Gerlitz for the Digital Methods Initiative. https://tools.digitalmethods.net/beta/trackerTracker/ [Accessed August 30, 2015].

Wayback Network Per Year. 2010. Developed by Erik Borra, Esther Weltevrede and Anne Helmond for the Digital Methods Initiative. https://tools.digitalmethods.net/beta/waybackNetworkPerYear [Accessed August 30, 2015].

Referenced projects

Below the research projects by the Digital Methods Initiative that are mentioned in this dissertation are listed. The project pages include further information such as query design, data sets, the different ways in which the data has been processed, techniques, methodology, high-resolution graphics and the project members.

Bekema, Vera, Liliana Bounegru, Andrea Fiore, Anne Helmond, Simon Marschall, Sabine Niederer, Bram Nijhof, Richard Rogers, and Elena Tiis. 2010. “Nationality of Issues. Rights Types.” Digital Methods Initiative Wiki. June 30. https://wiki.digitalmethods.net/Dmi/NationalityofIssues [Accessed August 30, 2015].

Borra, Erik, Taina Bucher, Carolin Gerlitz, Anne Helmond, and Esther Weltevrede. (2010). “One Day On the Internet Is Enough Aka Pace Online.” Digital Methods Initiative. October 4. https://wiki.digitalmethods.net/Dmi/OneDayOnTheInternetIsEnough. [Accessed August 30, 2015].

Dagdelen, Demet, Martin Feuz, Marije Rooze, Thomas Poell, and Esther Weltevrede. 2010. “Historical Controversies Now.” Historical Controversies Now. August 18. https://files.digitalmethods.net/var/historicalcontroversies [Accessed August 30, 2015].

DensityDesign, Digital Methods Initiative, Eurecat, and MédiaLab SciencesPo. 2015 “Contropedia.” Contropedia.net. http://contropedia.net/ [Accessed August 30, 2015].

Digital Methods Initiative. 2007. “Issue Image Analysis.” Digital Methods Initiative Wiki. July 15. https://wiki.digitalmethods.net/Dmi/IssueImageAnalysis [Accessed August 30, 2015].

Digital Methods Initiative. 2009. “Climate Change Skeptics.” Digital Methods Initiative Wiki. February 13. https://wiki.digitalmethods.net/Dmi/ClimateChangeSkeptics. [Accessed August 30, 2015].

Govcom.org. 2007. “The IssueDramaturg.” Issue Dramaturg. http://issuedramaturg.issuecrawler.net/about.html [Accessed August 30, 2015].

Rogers, Richard, Esther Weltevrede, Sabine Niederer, and Erik Borra. 2011. “National Web Studies: Mapping Iran Online.” Digitalmethods.net. http://mappingiranonline.digitalmethods.net [Accessed August 30, 2015].

Weltevrede, Esther, and Anne Helmond. 2012. “Where Do Bloggers Blog? Platform Transitions in the Historical Dutch Blogosphere.” http://dutchblogosphere.digitalmethods.net [Accessed August 30, 2015].

Introduction: A device-driven perspective to digital research

Chapter 1

This dissertation intends to contribute to current debates in digital research in two ways.[1] First, a disciplinary contribution to the areas of digital social research and software studies is made, by developing a 'device-driven' perspective which is attentive to the operational capacities of digital media and the ways in which these media can be made productive as sources of data for digital research. Second, a contribution to digital methods is made by introducing the notion of 'research affordances' in order to provide ways to operationalize the device-driven perspective and to discuss the role of digital media as sources of data and their implications for the research process. In what follows both contributions will be unfolded, but first some of the key debates in digital research are identified, starting with debates in digital social research.

Digital social research increasingly treats the web as collections of data (see for example Lazer et al. 2009; Pentland 2014; Manovich 2012b). The rise of social media platforms enable the collection of data sets anticipated to open up secrets about cultural life and society at large on an unprecedented scale. Other digital data sets, however, are also used for research, such as search engine query logs and historical versions of websites. Apart from opening up data for new lines of inquiry, the promises of digital data as a source for research include their being overabundant, easily accessible, and cheap (boyd and Crawford 2012). At the very least, the data produced on and by platforms and engines are an important practical input of various current programs of digital research, which have recently been tagged as big data, digital humanities, computational social science, database marketing and government surveillance. Data-driven research is practiced by a number of different disciplines, applying various methodologies and conferring a myriad of theoretical assumptions and expectations to these digital data sets.

One popular analytical approach to social media is the analysis of broad data samples in order to predict a range of social phenomena, such as elections (see for example Metaxas, Mustafaraj, and Gayo-Avello 2011; Gayo-Avello 2012). For a number of years now debates have developed around prediction-driven research and in probably the most-cited paper in the research area using Twitter data to predict the outcome of elections, the authors claim that Twitter data indeed are a good predictor for elections because the frequency of tweets mentioning a party or candidate closely followed the distribution of votes for the different parties in the 2009 German federal election (Tumasjan et al. 2010). However, the study prompted reactions with serious reservations on the accuracy of the method used; the main point is that the results risk to be modeled after the outcome of the election is known, besides related methodological issues, such as not including all parties in the analysis and varying results subject to

[1] The term 'digital research' will be used as shorthand for both digital social research, digital medium research and any combination thereof.

the time window chosen to analyze the tweets (Jungherr, Jürgens, and Schoen 2012). Digital data have brought novel opportunities and experiments for digital research, but the question is how and when digital methods are appropriate and relevant in terms of their ability to produce interesting and reliable findings.

My intention is not to provide a disciplinary overview of the emerging areas of digital social research.[2] Instead, my contribution specifically deals with the premise that currently too little attention is paid to how digital data enclose the analytical assumptions and expectations through which it has been informed by the digital medium and its cultures of use. It is often overlooked that the digital media producing and providing access to the data sets are methodological devices, too. This in turn challenges certain assumptions and expectations brought to the data by digital researchers. Moreover, focusing on web technologies and how they operate as devices in digital research methods has analytical value. They may function as an entry point into the connection between digital media, their cultures of use and the research apparatus including research questions, tools, data collection, analysis and findings. The various research programs working with digital (web) data are often referred to as 'data-driven' research (see for example Manovich 2012b, 12; Lazer et al. 2009, 2). I suggest to shift focus from data-driven digital research to what I call *device-driven* digital research; in so doing I address some of the opportunities and limitations of digital methods by focusing on the specific operations of the media as the source of our collections of data. I do so by connecting to the area of software studies and related fields in order to draw out the politics of the platforms and engines serving as sources of the data in web-based digital research. How, then, to deal with the epistemological challenges put forward by the use of digital media as sources of data?

In the edited volume *'Raw Data' is an Oxymoron* media historian Lisa Gitelman (2013) argues that we cannot take the 'data' in Big Data for granted. Drawing on a suggestion by Geoffrey Bowker, she reminds us that data are not simply 'raw', ready to be collected and analyzed for myriad purposes, but instead data are always 'cooked'; they are created and emerge from particular historical, economic, social and cultural circumstances. Data are often said to be 'collected', 'piled', or 'mined'; Lev Manovich points out that data do not just 'exist' but are always 'generated' (2013, 3). In that sense data always depend on knowledge production processes. Gitelman examines 'how different disciplines have imagined their objects and how different data sets harbor interpretative structures of their own imagining' (2013, 3). Precisely because

[2] For an overview and outlook of the field of digital social research programs see for example (Berry 2012; Lazer et al. 2009; Pentland 2014; Manovich 2012b; boyd and Crawford 2012; Crawford, Gray, and Miltner 2014; Borgman 2009; Svensson and Goldberg 2015; Gold 2012).

data are constructed, she argues, data 'need to be understood as framed and framing, understood, that is, according to the uses to which they are and can be put' (Gitelman 2013, 5). In order to grasp this framing, the multifaceted contexts in which data are created and the specific social and cultural phenomena encoded into the data must be further examined.

A growing number of empirical researchers endorses that digital data is not just 'there', ready to be collected and analyzed (Rogers 2013b; Puschmann and Burgess 2014; boyd and Crawford 2012; Rieder and Röhle 2012; Marres and Gerlitz 2015; Langlois and Elmer 2013). They point to a wide variety of issues ranging from the (unknown) demographics of digital data (boyd and Crawford 2012), to the baseline (Rogers 2013b), to trust in (or illusion of) machinic objectivity (Rieder and Röhle 2012) and the implications of commercial interests in the politics of knowledge (Langlois and Elmer 2013). Twitter data, for example, does not appear to be that suitable for predicting election results due to demographic representational issues (Google data might be more useful). Instead of predicting election results, however, Twitter data may be amenable to detecting signs of early buzz or warning as well as other societal monitoring. From a device-driven perspective this can be imagined, because the same media that are part of the communication and social coordination are also part of the methodology for these modes of research, creating alignment between the digital media operations and the research objective. Among others, I engage with the question what we can know about the data and how to know it. To derive findings from a data set, we need to understand what the data set entails and the conditions of how it was created (Uprichard 2013; Gerlitz and Rieder 2013; boyd and Crawford 2012). Moreover, the methodological reservations about the election prediction study are exemplary for how the Twitter platform pre-formats and activates its cultures of use and what can be studied when using Twitter as a data source. The epistemological question, then, concerns the reality the data capture (Savage 2009), since it is at the same time an expression of social reality and the product of a digital medium.

Empirical researchers have pointed to the epistemological ramifications of analytical tools on the definition of knowledge (Rogers 2013b; boyd and Crawford 2012; Rieder and Röhle 2012; Borra and Rieder 2014). Digital research and its computational tools open up new corpora, definitions of study objects, and methods to study them. Scholars have pointed out the tools at the researchers' disposal influence both research and its outcomes (see for example Uprichard, Burrows, and Byrne 2008), but the influence of digital media on shaping the data and research process is often overlooked. Digital platforms and engines have, however, the ability to become part and parcel of the researcher's tool kit, study object, research practice and method. *Wired*'s then editor-in-chief Chris Anderson posed the question, 'what can science learn from

Google' (2008)? The more critical question would be how popular digital media set the conditions of possibility for our procedures of knowledge production.

Whilst the term digital research might seem to reduce the importance of medium research, it emphasizes the digital, which fosters both medium functionalities, operations, cultures of use and social practices; this will be further specified as the argument unfolds. The digital media storing, sorting and delivering the data for research tend to be organized in ways that favor particular types of analysis, such as studies into people's social networks, website histories, and the spread, reach and currency of content at specific moments in time. Some of these privileged types of analysis closely follow the medium, such as research focusing on the history of single websites following how the Internet Archive's Wayback Machine provides historical versions of websites, which is searchable only on domain name (Rogers 2013b, 201; Stevenson 2010). Or consider how data sets collected from the Twitter Streaming Application Programming Interface (API), are presentist, lending themselves to study current events, as witnessed by research focused on elections, revolutions, disasters or trending topics (Bruns and Stieglitz 2012; Rogers 2013a).[3] In addition, digital data enable more complex and undetermined modes of analysis, such as the co-occurrence of terms in various digital platforms (Marres and Gerlitz 2015; Chapter 2), or hyperlink network analysis on Internet Archive data (see Chapter 4). When using digital data for research, it is important not only to understand the way digital media transform social and cultural phenomena, but also which interpretations of the data are appropriate in relation to the research question.

Device-driven research calls for a medium-specific perspective. By employing such a perspective I align myself with the Digital Methods Initiative program that advances medium-specific methods using web data (Rogers 2013b; Digital Methods Initiative n.d.). Medium-specificity refers to data, methods, and objects that are native to, not imported into the medium. In *Digital Methods* (2013b) Richard Rogers historically positions the digital methods research program by distinguishing between the natively digital and the digitized in relation to Internet research methods. With the notion of 'repurposing' he positions digital methods as research sensitive to the methods embedded in the medium, as opposed to applying existing methods from social sciences and humanities (2013b, 19). The debate outlined above about the relations between digital method and the digital media it collects data from is, for me, captured by the term 'repurposing' and engages with the tensions in the research process emerging

[3] Although the Streaming API is the easiest way to collect large collections of data, Twitter additionally offers a REST API and a Search API, each with their respective privileged modes of analysis and restrictions. See Chapter 2 for more on the specifics of these APIs; see also (Puschmann and Burgess 2014).

from working with digital (web) data. These are questions various digital methods researchers are currently engaging with, each in their own way (Rogers 2013b; Marres 2012; Borra and Rieder 2014; Marres and Gerlitz 2015; Niederer and Van Dijck 2010; Gerlitz and Helmond 2013; Langlois and Elmer 2013; Lury 2012; van der Velden 2014; Beer and Burrows 2013). I contribute to this repurposing debate by inquiring into the role digital media play in digital research by means of a 'digital device' perspective and its implications on the research process. In so doing I position myself in current debates by exploring digital media as devices in the research apparatus with the intention to connect digital methods for social research with medium research. This allows me to engage with some of the implicit and explicit critiques of digital methods research, which, as I will return to shortly, can be considered as concerns troubling digital culture and social life much more generally. But first I will address what I intend with the digital device.

The device as object and method

Although the term device in combination with method is potentially explosive as it connects to vast and ongoing academic debates about the role of devices in social research, my contribution is rather specific.[4] I bring a new media studies perspective to digital devices in relation to methods, seeking to pave the way in which digital media can explicitly be mobilized for knowledge production, through an inquiry into the politics and methodological ramifications inscribed into these digital media. I thereby contribute to the repurposing debate by examining questions that are currently being explored, such as the difference between 'social research' and 'medium research' and the alignment between medium and method. I am drawn to the term digital device because it allows me to bring together digital media as an *object* and at the same time as a *process* in the context of digital research. The term allows me to talk about the performative qualities of platforms and engines and about their affordances for research.

The term device has a number of everyday meanings and is used in myriad contexts, referring to objects designed for a specific purpose, such as gadgets, tools and bombs as well as methods, such as tactics, plans, stratagems and designs (OED Online 2015). Devices are, however, not mere tools; they are also complex and unstable arrangements bringing together a variety of people and objects with particular purposes. The

[4] See for example the work, informed by Science and Technology Studies (STS), about the performativity of methods in the enactment of the social (Law 2004; Ruppert 2007; Savage 2010; Law, Ruppert, and Savage 2011; Callon, Millo, and Muniesa 2007).

term digital device was the object of analysis in the 2013 special issue on 'The Devices' edited by John Law and Evelyn Ruppert (2013), where it was argued that digital devices enact the social in a similar way as performative devices in social sciences and economics (Law and Ruppert 2013). They emphasize that devices embed something into the very thing they seek to analyze. The special issue is meaningful in theorizing the role of the device in social research by demonstrating how a wide variety of devices used for knowing social and cultural lives are also part and parcel of the production and performance of contemporary cultural and social life.

Another recent collection of studies compiled around devices and how they are put to use in method is called, *Inventive Methods: The Happening of the Social* (2012). The editors Celia Lury and Nina Wakeford argue that 'inventiveness is not intrinsic to methods; it is rather something that emerges in relation to the purposes to which they are put' (2012, 2). In other words, they inquire into how the specific research objectives and other elements of the research apparatus are brought into dialog. In this vein, digital methods can be seen as inventive when they can 'change the problem they address', or put differently, when methods can 'self-organize in a (changing) relation to a (changing) context' (2012, 13). More recently, Noortje Marres and Carolin Gerlitz advanced the relation between digital media and social scientific method with so-called 'interface methods', which they understand as 'a process of the assembly of different components of the digital social research apparatus', focusing specifically on the similarities and differences between methods in social sciences and the methods of the medium from a digital sociology perspective (2015, 28).

The term device makes method more material, which is often overlooked (Lury and Wakeford 2012, 9). At the same time it invites a focus on the purposes inscribed into them—how devices articulate actions, act or make others act (Callon, Millo, and Muniesa 2007, 2). Simply stated, devices are often said to be *objects*; and secondly that they are performative and may contain *method*. With the digital device perspective advanced in relation to digital methods I relate the two propositions with examples to collapse the method-object division. Here the focus is on how digital media may function as devices in research, how they formalize and format relations and how they inform the (re)configuration of relations between object and method. A medium such as Facebook, for example, formats forms of sociality (such as 'friending' or 'liking') and at the same time allows for modes of analysis in academic research, government surveillance programs and database marketing (see for example Kramer, Guillory, and Hancock 2014; Payne et al. 2008; Holzner 2008). These research programs assemble analytical techniques, analysts, theoretical models, infrastructures and so forth to repurpose the data generated by—in this case—Facebook into something else.

The term digital device is productive to stress the performative role of digital media in both digital social life and digital methods. It allows me to move beyond the mere technical formalization of platforms and engines and to include the cultures of use and social practices. In other words, using data generated through Google or Facebook, the device perspective seeks to include into the analysis what I refer to as the 'device cultures' of these platforms and engines, taking into account the specific uses and practices inscribed in the data set under analysis. Device cultures can then be defined as the interaction between user and digital medium, the data collected, how they are analyzed, and ultimately the resulting recommendations (see Chapters 6 and 7). Web data, then, should not be regarded as generic, but as the result of specific device cultures. Such a perspective on digital media and its data does not solely focus on its software but takes into account the social arrangements and cultural practices that digital technologies incorporate and enable. The web consists of several types of digital media; here I focus on social media platforms, each hosting and formatting communication, dissemination and coordination in a specific way, and search engines, each having its own logic of ranking and scoring web content. I regard these platforms and engines as epistemological objects put to use as methodological devices in digital research.

The digital device can further be demarcated in terms of its scale. I focus on digital media repurposed as devices in the method. In social research the lineage of the device can be traced back to Foucault's notion of apparatus or dispositive, which is more extensive and is described as 'a thoroughly heterogeneous ensemble consisting of discourses, institutions, architectural forms, regulatory decisions, laws, administrative measures, scientific statements, philosophical, moral and philanthropic propositions,' he continues, 'the apparatus itself is the system of relations that can be established between these elements' (Foucault 1980, 194). Foucault furthermore suggests that the apparatus is always a collection of modifications in response to an urgent need or following a strategic function.[5] Compared to the larger notion of the apparatus, the act of repurposing digital media as devices in the method is more akin to the action

[5] Also see Giorgio Agamben who draws and expands on this already large class of Foucauldian apparatuses and calls an apparatus 'literally anything that has in some way the capacity to capture, orient, determine, intercept, model, control, or secure the gestures, behaviors, opinions, or discourses of living beings' (2009, 14).

of deploying *tactics* than to a *strategy*.[6] Tactics are isolated actions taking advantage of opportunities offered by a given strategic system; they are adaptive in exploiting strategic systems to generate novel and inventive outcomes. That resembles the relation between platforms and engines, and the scripts and scrapers built on top of them: researchers repurpose the analytical capacities of digital devices in their research design for novel and inventive outcomes. Although I consider the repurposing of the digital device as the tactical use of an operative digital media element in the method, relating the notion of the device to the lineage of the apparatus allows appreciating how a device does not operate in isolation and instead is always in relation and in configuration to other components.

In the context of digital research devices are epistemological objects capturing, processing, analyzing, ranking, recommending, formatting and aggregating data from the web. In the following chapters globally dominant digital media such as Google, Google Ad Planner, Wikipedia, Facebook, YouTube and Twitter are used as research devices, as well as media significant in very specific regions, such as Likekhor, Donbaleh, Sabzlink, Balatarin in Iran and the local domain Googles such as Google.nl, and in specific niches such as the Wayback Machine for accessing the historical web and Alexa for web analytics. Media that may have been significant at the time of the empirical case study but have changed their core business, such as blog search engine Technorati, or disappeared, such as Likekhor, Sabzlink, Google Reader, Google Blogsearch (Images) and Google News Images are also included.[7]

The notion of the device is thus productive in the context of digital research because it allows me to discuss the performative qualities of platforms and engines and their affordances for research. The alignment of research aim or question with an intentionally configured research device is key in the device-driven research perspective. The digital device finds itself therefore in between feature, which is often at the center of media operations, and apparatus, which encompasses the entire research assembly.

[6] This draws upon the work of the French scholar Michel de Certeau (1984) when he proposes that tactics are used to negotiate the strategies arranged by larger strategic systems such as organizations or institutions. It also resonates with the notion of tactical media, a critical practice of media making (Kluitenberg 2011). What ‘repurposing’ has in common with tactical media is that it seeks to ‘experiment with the medium’ (Raijmakers in Kluitenberg 2011, 19) for digital research but not necessarily for political activism.

[7] Without exception, the above-mentioned media have changed their algorithm, interface, business model and so on between the moment of the empirical research and the final written version of this work. In order to address how the volatility of digital media may implicate research, Chapter 5 follows the most important algorithm changes in the Google engine and how this prescribes the conditions of what can be known with the engine as a research device.

This is further specified by connecting to ideas developed in software studies, most notably concerning affordances and how they shape uses and practices (Stanfill 2015; Crawford and Gillespie 2014; Langlois, Elmer, et al. 2009; Bucher 2012a; Van Dijck and Poell 2013). With this software studies-informed perspective to digital research I approach these insights to inquire into how platforms and engines afford specific modes of research.

Digital media affordances

Approaching media from a device perspective allows approaching them in terms of their affordances and opportunities for research. The point of view of the spectator or user, which is generally considered as the preferred perspective in traditional media studies, is not the preferred vantage point to understand digital media. Instead, digital media are understood as devices—as the process of becoming activated. The term 'operationalisation' as advanced by literary scholar Franco Moretti (2013) is productive in relation to the role of the digital device in the method because it emphasizes how the analytical process is structured by the digital device. As such, operationalization is not just a one-directional move from method to implementation; Google as a device for research may therefore be put to varying uses in different research projects.

Drawing on the notion of 'sophistication' put forward by Matthew Fuller and Andrew Goffey (2012, 14), I advance the practice of imagining digital media in terms of their sensitivities and concerns as a prerequisite to use their analytical affordances in digital methods. The device perspective seeks to make sense of digital data through techniques and methods allowing for the development and deepening of tacit knowledge of the medium, with a specific interest in digital features, algorithms, functions, settings and so on. Tacit knowledge refers to the under-codified grasp that comes from the hands-on research practice with digital devices. The device perspective seeks to partially codify, or at least explicit, the kinds of hands-on skills required to empirically study digital devices by exploring how to research medium-specific objects, including websites, engines and platforms, but also how to deploy such objects for research. Digital media objects are operationalized as devices by, for example, treating default settings as affordances and features as content for analysis, by exploring the query and result pages, the relation between front-end and back-end and by examining APIs and streams. The approach to media I explore entails the cultivation of a certain medium-specific sensibility that seeks to appreciate the ways in which media objects can format and activate a diversity of practices and the way cultures of use find their conditions in the objects and forms of their media environments. The term device is thus understood as how things, technologies, techniques, cultures and practices become operative. I thereby connect to one of the core concerns of software studies, namely the question of the specific operations of software in digital culture and social life.

Here I return to software studies and related areas. Software studies is an interdisciplinary field studying software by using methods and theory from the digital humanities and from computational approaches to software, among others (Fuller 2003). Early calls to study software as a cultural practice appear in Lev Manovich' *The Language of New Media* (Manovich 2001) and Matthew Fuller's *Behind the Blip: Essays on the Culture of Software* (Fuller 2003).[8] The study of software and its cultural and social effects is interdisciplinary; a variety of approaches are followed, with an interest in the 'conditions of possibility' software establishes, referring to the operational capacities of how software acts and makes others act (Fuller 2003, 1). Related areas of study include critical code studies (see for example Marino 2006), which is more focused on the code rather than the program as a whole, and platform studies, which is focused more on data exchanges and the interoperability of platforms (see for example Gillespie 2010; Langlois and Elmer 2013; Helmond 2015). Software studies invites us to focus on the ways in which digital activity and practices are not simply human activity, but always 'negotiated through complex dynamics between software architectures and different categories of users (i.e. software engineers, citizen, activists, etc.)' (Langlois, McKelvey, et al. 2009). Methodologically, software studies emphasizes computer literacy and analyzes software sources and processes (see for example Goffey 2008; Gehl 2014), but also uses methods from existing fields such as social sciences (Kitchin and Dodge 2011, 255). In recent methodological contributions interface analysis is combined with the analysis of other material such as tech documentation in order to study the relationship between the interface and the code (Bucher 2012a; Gehl 2014). Despite its varied and rich set of approaches and methods, Manovich argues the field 'need[s] new methodologies' (Manovich 2013, 15). The objects and methods of study are varied and this is perhaps its strength; as software is increasingly diverse and all-pervasive, the field welcomes new methods on new objects (Gehl 2014, 8). In addition to contributing to social research with a software studies-informed device perspective, I develop a contribution to software studies that answers this call, by advancing digital methods to study computational media, which I refer to as 'medium research' and in which I am specifically drawn to Bucher's suggestion to develop methodological ways to let the software 'speak' (2012a, 74) (this will be taken up specifically in Chapters 3 and 5).

I connect to software studies as a field interested in the unstable and contingent nature of software, developing a critical understanding of media as devices themselves.

[8] For an overview of researchers that align themselves with software studies, see David Berry (2011b, 4–5) and Taina Bucher (2012a, 29–30).

The emphasis on software criticism rather than software use is particularly relevant when introducing digital media as devices into the method, that is, when the aim is to understand what these digital media do, and the politics and operations they entail, beyond providing data and functionality contained in them. The definition of software mobilized in software studies as a 'neighborhood of relations' is therefore not limited to only studying the material object of software (Bucher 2013; Mackenzie 2006, 169). The case studies in the following chapters give more empirical and theoretical accounts of the specific neighborhoods of relations, by exploring the specificity of digital media as devices in relation to the research apparatus.

By approaching media from a device perspective I also seek to connect insights from the field of platform studies with digital social research. I pursue the connection to the extent that the critical understanding of platforms leads to inventive methods. Tarleton Gillespie runs through the dictionary definitions of platform in "The Politics of Platforms" (2010), demonstrating how the platform allows the negotiation of different interests and purposes. Platforms carefully position themselves in relation to users, clients, advertisers and policymakers, strategically tailoring the platform's purposes to each in a different way, which sometimes leads to conflicts of interest. The term 'platform' is specific in the context of this argument as it reveals the contours of this discursive work, sometimes as technical 'platform', sometimes as 'platform' from which to speak, or as 'platform' of opportunity (Gillespie 2010). I argue, however, that researchers may critically intervene and repurpose platforms as research devices by taking into account the various purposes of the platform that are relevant to the direction of the research question.

The term 'affordances' is useful in the context of digital media as it calls attention to the object's qualities allowing specific actions. The term affordance is used in a variety of fields and was first introduced by the psychologist James Gibson in his article 'The Theory of Affordances' (1977), further developed by design theorist Donald Norman in *The Design of Everyday Things* (2002) and inferences were made from it in Human-Computer Interaction (HCI) (for example Preece et al. 1994). In *Behind the Blip* (Fuller 2003) Matthew Fuller calls for software critique drawing on the field of HCI because here the machinations of the computer are made available to the user. He writes, 'the way the computer makes available such use, and the assumptions made about what possible interactions might develop, are both fundamentally cultural' (Fuller 2003, 12). The term affordance encompasses the assumptions inscribed into digital media and the specific cultures of use privileged by its design. Affordances are also tied to those who engage with them and are different for different actors. Consider for example how the affordances of the empty profile, the progress bar and prompts on Facebook solicit data from users, which are used as new types of demographic categories affording advertisers to target their products in specific ways.

I introduce the term 'research affordances' of digital media, focusing specifically on the analytical affordances of platforms and engines as devices in digital research; they deal with the relation between objective, medium and method, and are specific to the actors and contexts of use. Considering the qualities of digital devices requires the overhaul of the theoretical assumptions of our methods, to make those assumptions explicit. More specifically, the research affordances of digital media require rethinking the technical forms and formats of the digital. Research affordances of digital devices mobilize the analytical capacities of the medium and their specific device cultures.

Furthermore, the notion of affordance allows examining the norms produced by, and politics of, digital media. Examining affordances illustrates their intended use by interpreting the medium's embedded assumptions about its purposes and how they can be properly used through the interface (Stanfill 2015; Van Dijck and Poell 2013). Digital media have 'politics', in that they can be seen as having 'powerful consequences for the social activities that happen with them, in the worlds imagined by them' (Gillespie 2003, 108). Affordances deal with the relation between software and user, and the tendencies built into the interfaces, in short, how action is structured by making 'some things more possible than others' (Stanfill 2015, 2). Not every user responds to the interface and its features in the same way, but the notion of affordances attends to the responses following the path of least resistance. Relatedly, the notion of 'default settings' is key when considering the preferred uses advanced through the design of media. In short, 'affordance' is a key concept when considering what software 'is suggestive of' (Bucher 2012a, 12). Accordingly, repurposing the research affordances of digital media is therefore a productive tactic, examining the possible uses of media by focusing on features, functions and categories of use and taking into account the specific cultures of use of the targeted data set.

Design affordances allow us to appreciate how the interface privileges some uses over other, but the interface of digital media should be recognized as embedded in software, and making use of 'protocological' objects regulating the conditions of possibility in a highly structured manner (Galloway 2004; Bucher 2012a). Digital media use technical standardizations and Internet protocols, such as those put in place by the World Wide Web Consortium (3WC), Internet Engineering Task Force and the Internet Society (ISOC). The hyperlink, for example, is standardized with the 'href' attribute (Helmond 2013).[9] As protocological assemblies, digital media provide the conditions of possibility for sharing content and digital data in an inclusive yet con-

[9] These technical formalizations and standardizations can however also be platform-specific, such as Facebook's Like Button. Officially called the Open Graph Protocol, despite being neither open nor a protocol, it uses W3C's Resource Description Framework (RDF), the foundational knowledge representation system of the Semantic Web (Halpin 2012).

trolled manner. As Alexander Galloway argues, protocol is not merely a technical specification but it must be understood as a management style governing the relations it contains (2004). Protocological objects can formalize, format and afford specific relations and data flows.

Phil Agre's concept 'grammars of action' (1994), which stems from computer system design, is particularly useful in further understanding the relation between code and the interface, and the way in which uses are pre-structured in digital media. The notion is part of the 'capture model', where the grammars of action and capture coincide and the pre-structuring not only concerns action, but also the data formats the grammars produce (Agre 1994). Where the notion of 'protocol' allows me to focus on how digital devices regulate relations and data flows, the notion of 'grammars' focuses on how activity and capture are prescribed into digital devices. The notions have in common that they are immanent to digital devices; there is no data exchange outside protocol, nor are there activities outside grammar. 'Grammars' specify the means by which actions might be composed by arranging actions into a sequence (for example, conditional and iterated sequences). Although grammars of action are not exclusive to the computer, digital media are readily understood as providing their different users with grammars of action, such as search bars, menu selections, like buttons, comment fields and so on. Digital media enact and capture ongoing activity through the imposition of a grammar of actions. The grammar becomes normative, because the users engaging in the articulated activity are somehow induced to organize their actions so that they are readily 'parseable' in terms of the grammar (Agre 1994). In addition to formalizing, formatting and affording action, the grammars of action thus tend to reduce the granularity of actions to make them readily available for capture and processing; for example, Facebook's like button is a reduction of comments such as 'awesome,' 'congrats,' or 'I took notice of this' into the grammar of 'liking' (Chan 2009).

The notion of device cultures calls attention how digital media's affordances become activated by their uses and practices in digital social life; it engages how users are imagined and prescribed into the interface through notions of affordances and grammars, and is attentive to the (un)intended uses and practices that take place on and with digital media. Device cultures can be related to the language metaphor of grammars of action, where they can be considered the spoken or lived action within media with a focus on the specific articulations of popularity, trends, and other frequency dynamics in the different media. The issue for digital research concerns how to interpret the specifics of the device cultures in the findings by recognizing activating features, signals and grammars, and by making the frequency dynamics productive in the analysis. From a device-driven research perspective, then, these device cultures can be used to focus the analysis and to find lively, trending, happening, popular, or other focal points relevant for the research objective at hand.

Drawing on developments in the above-mentioned areas I explore the analytic qualities of the digital, and in particular the research affordances of digital platforms and engines, by bringing together a software studies perspective with a device-driven social research approach under the umbrella of digital research. In that sense, this work is ‘new media research’ treating ‘media as software’ (Manovich 2013; Berry 2013) with the objective to empirically inquire into medium-specificity, including a focus on algorithms, scripts, buttons, links, and other natively digital objects with a performative function in the digital environment. Moreover, it focuses on the affordances of medium-specific features, algorithms and functions not only as research objects but also as methodological devices. In this way social issues and phenomena under study are always considered as mediated by the platforms and engines. The object of study is thus always partly social and partly medium. This medium-specific focus is interested in how the medium formats and shapes communicative acts and exchanges. The digital media device is introduced into research as an object of study and as operation in the method.

Having introduced what I intend with the digital device in digital research, I will now further engage with the role of the digital device in digital research by elaborating the notion of ‘repurposing’.

Repurposing and the research affordances of digital media

One of my central themes is that device-driven research allows collapsing the object-method distinction in digital research. The device perspective seeks to make the digital medium both part of the methodology and the object of study. This allows inquiring into how digital media format and formalize cultural, social and other relations, how we use these same devices in our methods, and what this means for our methods for knowing the cultural and the social. It also allows inquiring into the assumptions, logics, and purposes built into the digital devices to be made productive in the research process. In this context the notion of repurposing not only calls attention to the ‘re-distribution’ of digital research methods to the web (Marres 2012), but at once turns these media into objects and vehicles to study the social and cultural life supposedly enabled by them. Repurposing digital methods thus allows me to address some of the emerging considerations and limitations of digital methods, that is, to study the issues of digital research closely connected to issues troubling digital culture and social life more generally.

Moreover, the prefix ‘re’ in repurposing has a tradition in new media studies, highlighting the media studies discourse in device-driven digital research. Think for example of new media classics such as the term ‘remediation’ by Jay David Bolter and Richard Grusin (2000), which they considered a defining characteristic of new media, underlining how new media reproduce functions and characteristics of old media. And how the term ‘remix culture’ by Lawrence Lessig (2008) underlines the availability and relative simplicity of reusing digital material, challenging pre-digital copyright laws. The term repurposing advances a ‘research re’, one that stresses the availability and relative ease with which digital media allow us to reuse the analytical features and functions immanent in their core operations. Repurposing is a dynamic negotiation practice in which tools, practices, data and methods may be connected in diverse ways, enabling myriad analytic purposes.

The act of repurposing may be seen as related to the ‘purposes’ of the digital medium. The objectives and purposes of any digital medium, however, are diverse and volatile, since settings and functions evolve and negotiate the various interests of changing user groups, advertisers, third parties and developers (Langlois and Elmer 2013; Gillespie 2010). Whereas some digital media have a very explicit purpose, such as the Wikipedia page entitled ‘Purpose’ addressing the platform’s objectives (Wikipedia contributors 2014b; also see Chapter 6), most digital media are less explicit. Their purposes and objectives may be read from corporate slogans, advertisers documentation, technical documentation, and changes in settings, algorithms and functions. As will be elaborated upon in Chapter 2, the medium’s assumptions can be considered as ‘alien’ from a social research perspective, because the methods built into popular media have a variety of disciplinary backgrounds and serve the manifold objectives of the platforms rather than those of social researcher. By building scripts and scrapers on top of the digital media, researchers negotiate with the assumptions and purposes inscribed into these media and appropriate and recontextualize them in their research design (see Chapters 2 and 4; also see Marres and Gerlitz 2015).

In addition to how digital methods repurpose the research affordances of digital media, I seek to mobilize the device by situating digital methods as ‘thick’ methods at the intersection of interpretative and quantitative research. With digital data entering the humanities and social science, the old debate about the qualitative interpreting stories versus the quantitative producing facts is being renewed, too (Venturini and Latour 2010). While methods in the humanities and social science traditionally tend to follow a deductive, top-down, or theory-driven approach, methods in computer science tend to follow an inductive, bottom-up, or data-driven approach (Rieder and

Röhle 2012, 7). Digital media suggest specific ways to view and interpret the data at hand. According to Bernhard Rieder and Theo Röhle, digital technologies change the way in which scholars work with their material, how they approach it and interact with it. The question is how to address the methodological assumptions, purposes and models embedded in the devices informing our methods.

Digital methods can be considered thick methods because they draw together qualitative and quantitative, numerical and interpretative, universal and specific traditions. The digital media that host, sort, archive, process and serve the data come with built-in assumptions, imaginations, and purposes. These invested assumptions have different origins, for example from the computer sciences as the software companies building these devices use computational models and theories, and social science as algorithmic media often incorporate theories and models of for instance (social) network and textual analysis (Marres 2012; Brin and Page 1998; Page et al. 1999; Rieder 2012). The digital methods researcher's analytic work on the other hand may be said to provide what Clifford Geertz and others have called 'thick descriptions', describing both an empirical observation and its wider implications (1973). This is comparable to how Ganaele Langlois and Greg Elmer use the term 'thick data' (as opposed to 'big data'), underlining how digital objects allow to trace the articulations of technical, corporate and media logics (2013). Similarly, thick method introduces a different outlook to methods than is provided for instance by increasingly dominant big data, which tends to be 'thin' and reductionist. With thick methods, on the other hand, researchers consider the complex methodological processes underlying the analysis of the data, method and analytics built into media, to be key to their research.

Thick methods are also a way to cope with the commercial imperatives shaping the formation of digital media and their rhetorical presentation (see for example Gillespie 2010). It has been argued that we are entering a field of social research operating on the intersection between 'commercial' and academic research (Langlois and Elmer 2013; Burrows and Gane 2006). Talking about 'repurposing' from a social research perspective, David Beer advanced that these social media platforms can be imported into analysis 'in order to explore new ways of seeing, this would be to reshape the purposes of the software to our ends' (Beer 2012, 3). He continues by stressing the necessity of considering the divergent interests of commercial and academic research and to explore 'how we might reshape the use of such software to suit the agenda of a more critical form of social research' (Beer 2012, 3). Repurposing as a specific response to the alien commercial assumptions in digital media focuses on embracing and making productive use of these ambiguities.

In order to do so, the boundaries between the medium's digital methods and the researcher's digital methods should be considered flexible and intertwining. Digital researchers have proposed notions to draw attention to the instability of methods like ‘live methods’ (Lury 2012; Back and Puwar 2012), ‘inventive methods’ (Lury and Wakeford 2012) and ‘interface methods’ (Marres and Gerlitz 2015). One of the key themes in the repurposing debate is the extent to which digital methods researchers study the medium or the social (Rogers 2013b; Marres and Gerlitz 2015). The difference between medium research and social research is hence probably best understood as a difference in degree: in some cases, digital devices play an ostensibly large role in the structuring of data, while in other cases we can delineate a discernable empirical object, which, although the conditions of possibility are demarcated by the device, is not really reducible to the medium-architecture enabling it (this will be one of the central explorations in Chapter 2). Both the digital media and the digital method can thus be considered serving a multiplicity of analytic and normative, not necessarily transparent, purposes. As mentioned, digital methods are inventive when they ‘change the problem to which they are addressed’ (Lury and Wakeford 2012, 13); the digital method is assembled through its specific configuration of the device’s purposes, device cultures and aligned with the specific research questions and objectives. Digital methods are thus emerging as ‘thick’ methods at the interplay of an empirical engagement with methods in digital cultures and the wider research apparatus. In the following chapters, such device-driven digital research is further developed and explicated through empirical and theoretical means.

Issues in device-driven digital research

On the basis of six case studies I empirically and conceptually explore how the data outputted by a variety of devices can be used for digital research. I do so by engaging with research affordances and with three contemporary issues in digital research. The first engages with the notion of medium dependency and the way in which digital media and digital research are related. This is developed by inquiring into ‘social research’ in Chapter 2 and ‘medium research’ in Chapter 3. The second issue addresses the volatility of method in digital research, by engaging the flexibility and indeterminacy of the various components of the methodological apparatus to study the evolution of software in the Dutch blogosphere in Chapter 4, and by engaging with the evolution of the Google algorithms and how this has afforded different modes of research over the years in Chapter 5. The final consideration focuses on the notion of device cultures, introducing media uses, practices and their frequency dynamics as key components for digital research. Chapter 6 empirically explores how Wikipedia functions as a key device culture in contemporary digital culture and how it’s dynamics may be repurposed for research; Chapter 7 offers a comparative device culture analysis as a productive way to sample relevant URLs for censorship research. Each of the case studies addresses key steps in the configuration of the research device, including query design, demarcation, indicators, and analysis. They all lead to what I term 'modes of research' that bring together the operations of devices and research objectives productively. Let me further detail the key contributions of each chapter.

Chapter 2, SCRAPERS AND APIS AS RESEARCH DEVICES, places the use of digital media as source of data into a social research perspective and focuses on the available key data collection techniques as devices in the research process. I advance the device-driven approach and contribute to social research by focusing on the affordances of the medium under study and by considering how they may (or may not) be repurposed for digital social research. Such repurposing of a medium is brought about by connecting how something is captured with particular research interests and how this allows for particular modes of research. By looking at the tool building techniques for extracting data from platforms and engines—scraping and calling APIs, the chapter seeks to draw out the intricate connections between collection and analysis, and between digital media and research tools. The materials studied include API documentation of specific platforms as well as the documentation of scrapers. Methodologically, this chapter seeks to draw connections between different steps in the research process; data collection and its relation to query design, selection, data cleaning, and analysis is emphasized. The chapter engages with the repurposing debate by empirically and conceptually exploring the difference between researching the medium and researching the social. The case study is used to differentiate between 'liveliness' and 'liveness' in real-time research, which, in brief, is the difference between measuring frequency versus measuring engagement.

Chapter 3, THE POLITICS OF REAL-TIME, takes a medium research perspective which is distinct from the social research perspective in the previous chapter; I introduce it as a form of software studies by examining what digital methods can add to the area of study. I investigate what various platforms and engines capture, format, and formalize and how they present, rank, order, and prioritize this as data. Methodologically, this chapter contributes to software studies by advancing comparative device-driven data analysis to empirically study the construction of different realtimes by digital media. In so doing the platforms and engines are repurposed as devices for software studies. Different kinds of materials are studied, ranging from patents, developer blogs, help pages, information gathered by watchdogs and trade press, to interfaces and (default) settings, as well as particular cultures of use. I contribute to software studies by repurposing digital devices in specific configurations and by using the data outputted as material for what I term 'medium research'. I do so by engaging empirically with the politics of realtime in various digital media. The case study suggests that realtime cannot be accounted for as a universal temporal frame in which events happen; the chapter explores the making of realtime in different platforms. Based on an empirical study investigating the pace at which various digital media produce new content, the different rhythms, patterns or tempos created by the interplay of devices, users' web activities and issues are traced. Distinct forms of 'realtimeness' emerge, which are not external from, but specific to devices, organized through socio-technical arrangements and practices of use. Realtimeness thus unflattens more general accounts of the realtime web and research, and draws attention to the operations built into specific platform temporalities and the political economies of making realtime.

Chapter 4, CONJURING UP A PAST STATE OF THE WEB, engages with the volatility of method in digital research, by focusing on how specific configurations of digital devices, research aims and modes of analysis may affect the behavior of their components. It inquires into the procedures by which the Wayback Machine selects and reformats digital data, and looks into the extent to which it allows for other types of research besides the single site history—which, as mentioned, is what the interface of the Wayback Machine privileges (Rogers 2013b). The chapter proposes a methodology to repurpose the Wayback Machine so as to trace and map transitions in linking technologies and practices in a blogosphere from 1999 until 2009. By using traces of technical markers, techniques and methods for historical network analysis are introduced, and the temporal dynamics of the Dutch blogosphere are analyzed. Such an approach enables the study of the rise and fall of blog platforms and social media platforms within the blogosphere and it enables the investigation of local blog cultures. The use of archived data for research, however, also brings to light some of its limitations, such as incomplete data sets and the loss of contextual data.

Chapter 5, GOOGLE ALGORITHMS AND THE VOLATILITY OF METHODS, also engages with the volatility of methods by discussing the extent to which the update culture of key web technologies leads to changing research affordances. I call attention to how device-driven research makes online platforms and engines—Google in this case—into a study object as well as a process structuring and changing the digital research enabled by it. By studying a variety of materials—ranging from patents and trade press to watchdogs—the chapter inquires into Google's capturing and organizing logics. By regarding these logics as the conditions of possibility for what can be known with them, they are made empirically researchable. I present a periodization and overview of Google's key algorithm changes and connect them to the modes of research they afford or discontinue, such as ranking research, national web research, realtime research, personalization research and source-distance research. As such, this chapter lays the foundations for the digital methods' notion of 'search as research' (Rogers 2013b) and further develops the digital methods approach to software studies. This approach offers an opportunity to discuss some of the limits of digital research and their resemblance to issues such as volatility, but also related issues troubling digital social life, such as the black-boxing of knowledge technologies. The epistemic trouble explored by digital research is then not just a problem of social research, but rather that of the many social practices involving the collection, management and analysis of digital social data.

Chapter 6, WIKIPEDIA AND THE VALUE OF DISPUTE, introduces the notion of 'device cultures' in digital research to draw attention to how cultures of use and social practices are key components in shaping the purposes of digital devices and how they may be productively repurposed for digital research. The chapter inquires into Wikipedia's encyclopedic apparatus—a bureaucratic apparatus of policies, guidelines and essays—and connects it to the processes of knowledge production through its content management system. To maintain its 'encyclopedianess' the platform has mechanisms in place with which consensus is designed, such as the core content policies' 'neutral point of view', 'verifiability', and 'no original research'. The back-end of an article, its edit history and talk pages document the work involved in reaching consensus. In this chapter it is argued that Wikipedia's socio-technical apparatus may thus be mobilized to trace and map controversies, since its prime aim is to defuse controversies. In so doing, the chapter is a contribution to controversy mapping (Latour 2005; Venturini 2010) and the related area of issue mapping (see for example, Rogers, Sánchez-Querubín, and Kil 2015; Marres 2015; Marres 2005). In order to underpin the idea of repurposing Wikipedia for controversy research the chapter debates the quality of the online encyclopedia's content. Previously this was studied by considering the collaborative processes in an article's edit history and talk pages, but the processes of dispute with which the dynamics of knowledge production involving societal issues can be studied received little attention. The method then focuses on operationalizing the appropriate natively digital objects and indicators for the research purpose at hand.

Chapter 7, THE NATIONAL WEB ACROSS DEVICES, offers a comparative device cultures approach to study the Iranian national web across a variety of its significant digital media. It offers an approach to conceptualize, demarcate, and analyze a national web. Instead of a principled definition of the types of websites to be included in a national web (as an archivist would do), the proposed method is adaptive and uses web devices providing (ranked) lists of URLs relevant to a particular country. The digital media are analyzed in terms of their device cultures and the latter's specific analytic value. These resulting lists of URLs are subsequently studied on the basis of certain common measures (such as responsiveness and page age), and repurposed to study censorship in the areas of the national web under investigation. The method focuses on demarcation by means of different devices, rendering them comparative and developing a chain of metrics to make claims about both the liveliness of a web and the extent to which it is censored. The chapter contributes to the field of Internet censorship research by developing ways to study its effectiveness and circumvention.

The case studies introduced may thus also be read as contributions to a set of relevant debates in contemporary digital culture. The digital media that are repurposed in device-driven research are also part and parcel of a more general digital cultural and social life, which is an argument I will further develop throughout this dissertation. The different chapters introducing epistemic issues related to digital research can also be read as broader issues in digital culture. For example, the issue of medium dependency in digital social research, which is connected to the decline of scraping and the rise of APIs, is also a relevant debate in digital culture as it exemplifies how platforms increasingly control derivatives (Bucher 2013). Chapter 2 advances this as an issue in digital culture; it engages with the difference between scraping and calling APIs, which each sets the conditions of re-using data and functionality differently. Chapter 3 engages with the debate about realtime as a technological imperative, which signals a 'fundamental shift from the static archive toward "flow"' (Lovink 2012, 11). I do so through an empirical inquiry into the construction of realtime in various platforms and by discussing sticky content in the realtime streams as a key mechanism to focus on content. The issue of volatility is also a concern that resonates well beyond the confines of digital research and affects digital cultural and social life, with notions such as 'perpetual beta' signaling a digital culture driving on 'continually updated services' (O'Reilly 2007, 14&1). The notion of volatility is also relevant in two competing narratives concerning the digital, one stressing its enduring ephemerality (Chun 2008) and the other its incapacity to forget (Mayer-Schönberger 2011). In Chapter 4 I investigate the decline of the blogosphere and the rise of social media platforms. The case narrates the evolution of the Dutch blogosphere as a DIY software development to off-the-shelf platform software. In Chapter 5, I engage with the debate about the volatility of web technologies, which focuses on how the update culture of key web technologies affects changing affordances for use as well as affecting changes in key knowledge logics in digital culture. Chapter 6 engages with the instability of truth claims and knowledge production, which is arguably enhanced by digital media (Latour 2007) and contributes to efforts developing different ways to evaluate the quality of controversies (Latour 2005; Venturini 2010). The chapter in particular closely investigates the mechanisms of Wikipedia to ensure the encyclopedic quality of the knowledge created on the platform. Another debate in digital culture, which is taken up in Chapter 7, is the rise of national webs and the related rise of jurisdiction and censorship (Deibert et al. 2010; Goldsmith and Wu 2006). The contribution in the chapter focuses on the issue of drawing online boundaries.

In the concluding chapter I seek to draw the findings of the specific case studies together in order to arrive at generally applicable observations. I conclude by proposing to slightly shift the focus in the repurposing debate to the digital media as devices in the research process, thereby shifting the focus to what may be considered a 'good' digital device and when. By focusing on the device I do not claim that the quality of the device is an inherent property, but rather I call attention to how we can consider the quality of the assembly of all components in the research apparatus by focusing on the digital device. The quality of the device thus depends on its specific context and I will consider the methodological configuration in terms of the three concerns outlined above: medium dependency in social or medium research, handling volatility productively, and making the specific dynamics of device cultures productive in the findings. In other words, I will empirically and conceptually explore how to align the device productively with the research objective and how it allows the medium to participate in identifying key issues and questions. I repurpose the operations embedded in digital media to address the issues of medium dependency, volatility and device cultures that concern digital methods. This is what I ultimately mean with 'repurposing digital methods'.

Scrapers and APIs as devices in digital research

Chapter 2

IN THE INTRODUCTORY CHAPTER digital research is positioned as operating at the intersection of medium research and digital social research. In this chapter I approach digital media from the digital social research perspective. I do so by approaching digital media as sources of data and by inquiring into the formats and analytics inscribed into the captured data, brought into the research process by two key data collection techniques, scraping and calling Application Programming Interfaces (APIs). Moreover, I discuss scrapers and APIs as analytical devices in digital research methods and I take up the repurposing debate by engaging the question how we distinguish between medium research and social research.

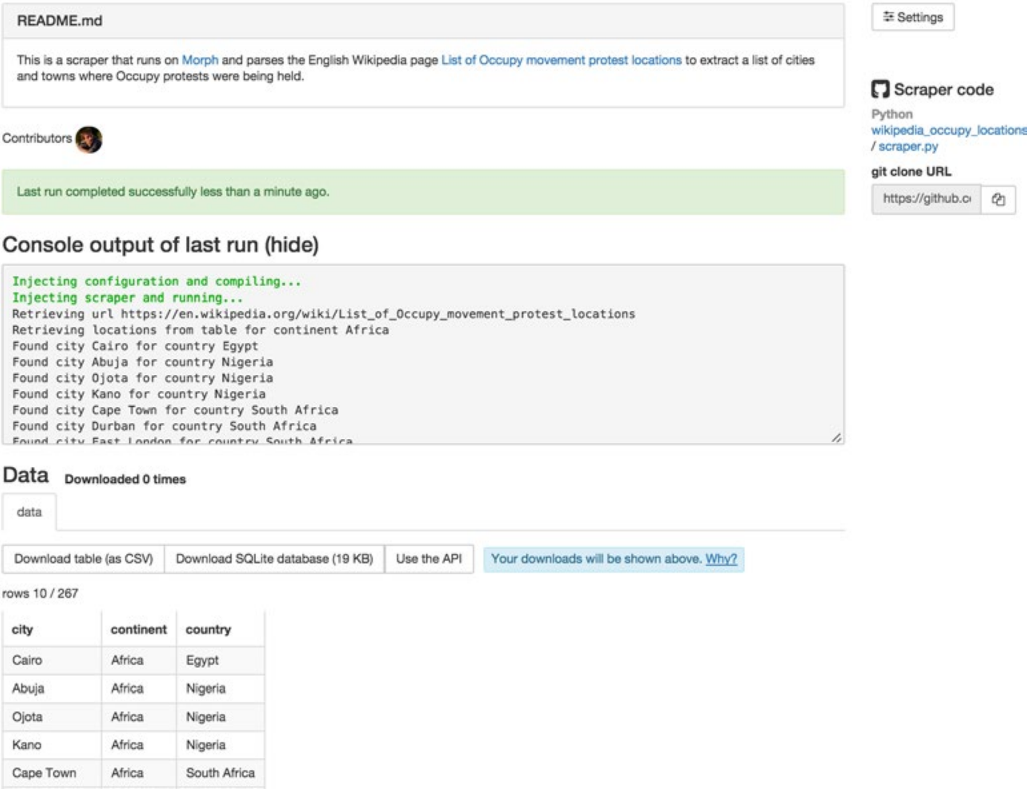
Scraping and calling APIs are prominent techniques for the automated collection of digital data and distinctive practices associated with current forms of digital social research, which are marked by the rise of the Internet and the ubiquity of digital data in digital culture and social life. Both are data collection techniques, but whereas screen scraping is an early technique to collect data from websites and social media platforms, APIs gained prominence with notions of the 'web as platform'; they are the preferred industry techniques enabling platform owners to set the conditions for access to the data (O'Reilly 2007; Helmond 2015). Scrapers and APIs make it possible to automatically download data from the web, and to capture some of the large quantities of data about social life available on digital media like Google, Twitter, Facebook and Wikipedia. Both are widely seen to offer new opportunities for digital social research too by enabling the development of new ways of collecting, analyzing, and visualizing social data. These opportunities have given rise to various declarations on the future of digital research labeled as the computational turn in social research, digital humanities and big data research (see for example Lazer et al. 2009; Bollier 2010; Berry 2012). As an initial entry point into this area I will focus on the relatively mundane practices and devices of scraping and calling APIs themselves. As techniques of digital data extraction both seem of special interest, because they are an important part of what makes digital social research practically possible. Moreover, they make use of the analytical affordances of digital media and make the tactic of repurposing practically possible.

This chapter explores the capacity of digital data collection to transform social research and to reconfigure the relations between research subjects, objects, methods and techniques. The devices of scraping and APIs are examined, as well as the kinds of research they enable. Such an approach allows digital research to be examined from the standpoint of its apparatus (Back 2010; Savage, Ruppert, and Law 2010). It enables a focus on the concrete and everyday techniques and practices of digital data capture. Such an approach makes it possible to adopt notions about the device as discussed in the introductory Chapter 1, and considers these in the context of digital research by introducing scraping and calling APIs as prominent data collection techniques.

In what follows, the argument is examined that device-driven research offers ways to unfold the method-object distinction as was introduced in the introductory Chapter 1, since scraping and APIs may inform the development of 'live' or 'real-time' modes of research (Lury 2012; Back and Puwar 2012; Elmer 2013). It may be argued that scraping and APIs introduce questions of medium dependency into the social research practice as the distinction between medium and social research becomes less apparent. Scrapers and APIs as data collection techniques were mainly developed outside social research: their recent popularity is closely associated with the rise of the 'real-time web,' which is, as will be explained, an industry term. Using scrapers and APIs to retrieve data within digital research, however, arguably enable a distinctive approach to knowledge-making across digital culture and social life. This approach is pre-occupied with monitoring happening or trending content, which is afforded by the streams and APIs of realtime media. This circumstance defines both the key challenge and opportunity of scraping and APIs for social research: if scraping and APIs are already deployed on the web, what does digital methods contribute by using scraping and APIs in a social research context? What might be the additional or alternative purposes of scrapers and APIs in digital methods?

In the second half of the chapter, some of the distinctive applications of scraping and APIs developed in recent digital research are presented. The most interesting of these applications, in the context of this work, are fully aware of the fact that scrapers and APIs are native to the web. The ways in which scraping and calling APIs entail web-based assumptions are interesting, as they are more than just techniques and bring categories into the research practice: scrapers and APIs present already formatted data for social research. It will be argued that this allows a distinctive approach to social research, approaching the formatting of digital data as a source of social insight as one of the core principles of digital research.

This argument is developed through a discussion of diverse empirical materials. In keeping with the understanding of scrapers and APIs as multifarious devices, an overview of these techniques is assembled from a variety of sources: from scraper and API manuals to popular scraper- and API-based research, to technical scientific articles on the subject. Once the features of scraping and calling APIs are established as research techniques and practices, a case study is presented, in which Noortje Marres and I scraped Google and called Twitter's API for particular keywords, 'austerity' and 'crisis'. This case study gives an insight in the distinctive form of digital social research enabled by scrapers and APIs, which is further developed with case studies throughout the other chapters. Provided that one learns how to take advantage of the data formats that scrapers and APIs open up for analysis, they represent powerful instru-



README.md

This is a scraper that runs on Morph and parses the English Wikipedia page List of Occupy movement protest locations to extract a list of cities and towns where Occupy protests were being held.

Contributors

Last run completed successfully less than a minute ago.

Console output of last run (hide)

```
Injecting configuration and compiling...
Injecting scraper and running...
Retrieving url https://en.wikipedia.org/wiki/List_of_Occupy_movement_protest_locations
Retrieving locations from table for continent Africa
Found city Cairo for country Egypt
Found city Abuja for country Nigeria
Found city Ojota for country Nigeria
Found city Kano for country Nigeria
Found city Cape Town for country South Africa
Found city Durban for country South Africa
Found city East London for country South Africa
```

Data Downloaded 0 times

data
Download table (as CSV)
Download SQLite database (19 KB)
Use the API
Your downloads will be shown above. Why?
Settings
Scraper code
Python
wikipedia_occupy_locations / scraper.py
git clone URL
https://github.c

rows 10 / 267

city	continent	country
Cairo	Africa	Egypt
Abuja	Africa	Nigeria
Ojota	Africa	Nigeria
Kano	Africa	Nigeria
Cape Town	Africa	South Africa

Figure 1: Partial screenshot of the live view of a morph.io web scraper that retrieves continents, countries, and cities from the Wikipedia page 'List of Occupy movement protest locations'. Shown are the console output and a partial view of the retrieved data. *Source: Borra 2015*.

ments for social research, extending the grasp precisely beyond what is happening 'right now'. But first the scrapers and APIs themselves are discussed.

Scraping and calling APIs as distinct data collection techniques

To begin with an example, Figure 1 shows a scraper at work. This screenshot captures a live view of scraping: it shows the moment when the scraper script is running and extracting data from the web. The particular scraper in the screenshot is extracting information from the online encyclopedia Wikipedia. It is querying a specific page, one called 'List of Occupy movement protest locations', in order to extract a list of cities and towns where Occupy protests were being held at that time. This sounds relatively straightforward, and indeed it is, insofar as this scraper is scraping only this one page, the title of which already describes the type of data it is scraping for: a list of cities and towns. However, the example also raises some less straightforward questions. Scraping is usually described as a technique for data collection, but is this scraper not analyzing data as well? After all, this script is not indiscriminately extracting data from the web, but only information on a particular page that fits a particular category: 'Occupy movement protest locations'. As the data is extracted from a web page that is made manually by various Wikipedia editors, the scraper has to identify where exactly on the page the various cities, countries, and continents are listed. The Wikipedia page has separate subsections per continent, after which various tables list in one column country or region, sometimes also called province or territory; in another column city or district, sometimes also called cities; one column containing the date at which the protests began; et cetera. The scraper of the Wikipedia page, thus not only has to indicate which data on the page is of interest, but in order to make useful use of it, it also has to name it (i.e. countries, regions, provinces, and territories will all be named 'country' and city, district, and cities will all be named 'city'). In the process of extracting data, then, it also identifies and formats particular data of interest.

Precisely this capacity of scraping to extract and construct 'structured information' from sources is highlighted in the formal definition of scraping to be found in technical literature on informational retrieval. While screen scraping is typically targeting digital data, automatic data capture has a much broader application, and a much longer history, going back to at least the 1970s. Information extraction is generally defined as 'the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources' (Sarawagi 2007). Arguably, it is precisely when information retrieval relies on information structures to extract data, that it has acquired special saliency in the current context of digital culture and social life.[10] In this context, the issue of information overload and the availability of overwhelming quantities of data are widely regarded as a central challenge and opportunity (Moens 2006). Information extraction promises a medium-specific solution to this problematic, in that it offers a way to extract—quite literally—relevant information from the data deluges made available by digital and networked media. On the web, information extraction addresses the

[10] The relevance of the semantic web should be noted in this context as the sematic web is the specification of content structures (for example title, summary, rating) and annotation structures which assists in data disclosure and analysis (Berners-Lee et al. 2001). In this chapter, however, emergent structures in web data are focused upon, rather than on efforts to implement them along such systemic lines.

problem of relevance by deriving structured data out of the heterogeneous expanse of digital information environments. It offers a solution by providing a way of extracting specific fields or data elements from pages on the web and other Internet sources, turning them into usable, well-ordered data sets.

As scraping is applied digitally, it becomes possible to gather data from multiple locations, putting the extracted data to new uses, thereby enabling the repurposing of digital data, which is one of the core concepts explored here. One can scrape digital media for images, or location data (geographic coordinates, town, region, country) or for text, which can then serve as a data set for research. Insofar as scraping relies on information structures, the technique is easiest to use if the targeted content itself has structure, for instance, if it takes the form of a table, or is formatted as 'tweets'. In another sense, however, scraping treats the web blindly, as a limitless expanse from which only specific elements need to be brought in. Scraping is however not limitless, as it needs to be tailor made for many different sites (for example, Wordpress formats differently from Blogger, not to mention BBC, CNN, and so on). A custom scraper thus needs to be devised for each content generator and tweaked for the specifics of the site. The resulting data, however, may be made comparable through formatting, for example by annotating scraped things as titles, or authors, or by putting all dates or locations in the same format.

Calling APIs, and more specifically web APIs, is a similarly prominent digital data collection technique, which, as mentioned, can be situated historically with the rise of the 'web as platform' (O'Reilly 2007; Helmond 2015), and to some extent succeeded screen scraping as a prominent collection technique, since APIs are generally easier to use.[11] Conceptually the data collected through an API shares core characteristics that are relevant for the development of the argument in this chapter—that the collection techniques make the repurposing tactic practically possible, but the API significantly differs from screen scraping. APIs are part of the computational definition of 'platforms' (Gillespie 2010) and in its most basic definition allow for *interoperability* between different systems, or put differently, allow a product or service to talk to another product or service (Bodle 2011; Bechmann 2013). More specifically web APIs are protocologial objects that regulate the access to data and functionality in a controlled manner via Hypertext Transfer Protocol (HTTP) (Bucher 2013). Web APIs 'provide information to third-party applications through "calls", a technique of retrieving data on a server in the background, without disrupting the display and function of a web

[11] For the origin of the notion of the 'web as platform', also see ProgrammableWeb.com; founder John Musser on the about page in 2005: 'ProgrammableWeb is a web-as-platform reference site and blog delivering news, information and resources for developing applications using the Web 2.0 APIs' (Musser, John 2005).

page' (Bodle 2011). Initially the notion of an API was developed in a programming context and it was meant to provide a fixed and structured interface to the functionalities or data of a program. The API formats software components as building blocks in terms of their operations, inputs, outputs and underlying types. In so doing, APIs define operations and functionalities independent of their specific implementations, which allows for variations in the uses of the software components. APIs can thus be seen as providing building blocks that can put together for specific uses. An API allows data to be shared within a software program, but also across applications. When the data and functionalities are made accessible via a network data and functionalities can also be shared externally with partners and the wider public. Originally APIs were used to integrate distributed systems by API-pioneering businesses like SalesForce, Amazon and Ebay and around 2006 the use of APIs on the web had evolved into new types of ecologies, making the web more 'social' which became clear when social media platforms such as Del.icio.us, Flickr, Facebook and Twitter started providing web APIs.[12] The web API also allowed platforms to tap into the creativity of third party developers (see for example Bucher 2013 on this point).[13]

APIs are an interesting research collection device because they provide the conditions of possibility for sharing content and collecting data, hence inscribing research affordances into the data circulation device. Moreover, as a protocological object the API allows very structured and controlled access to a platform. Twitter, for instance offers two different APIs based on the HTTP standard: the REST and the streaming API; both APIs have their own limitations. For instance, the REST API provides access to a variety of core functionalities and data points, such as recent tweets (up to a week), profile information and a user's followers, but the API has a small rate limit.[14] The streaming API, on the other hand, allows access to data in near realtime with a significantly higher rate limit, but does not provide access to as many data points;

[12] Although APIs were already around in the 1980's in software and hardware development, the 'web API' discussed here is just 10 years old. APIs are used for commerce, payments, social, cloud computing, mobile phones and tablets, and much more. On the API Evangelist technology professional and API enthusiast Kin Lane has collected a history of important pioneers and developments in the development of the API (Lane, Kin n.d.). Henceforth, when I mention APIs I mean APIs made accessible via the web.

[13] Following this definition the Google products generally lack a good API—except Google Maps—and cannot be considered a platform, therefore missing out on harnessing the collective intelligence of developers. This point was made by a Googler posting an internal rant about this on Google Plus but forgetting to uncheck the 'public' box (Yegge, Steve 2012).

[14] The developer documentation gives more details on the limitations for each of the APIs (Twitter Developers n.d.).

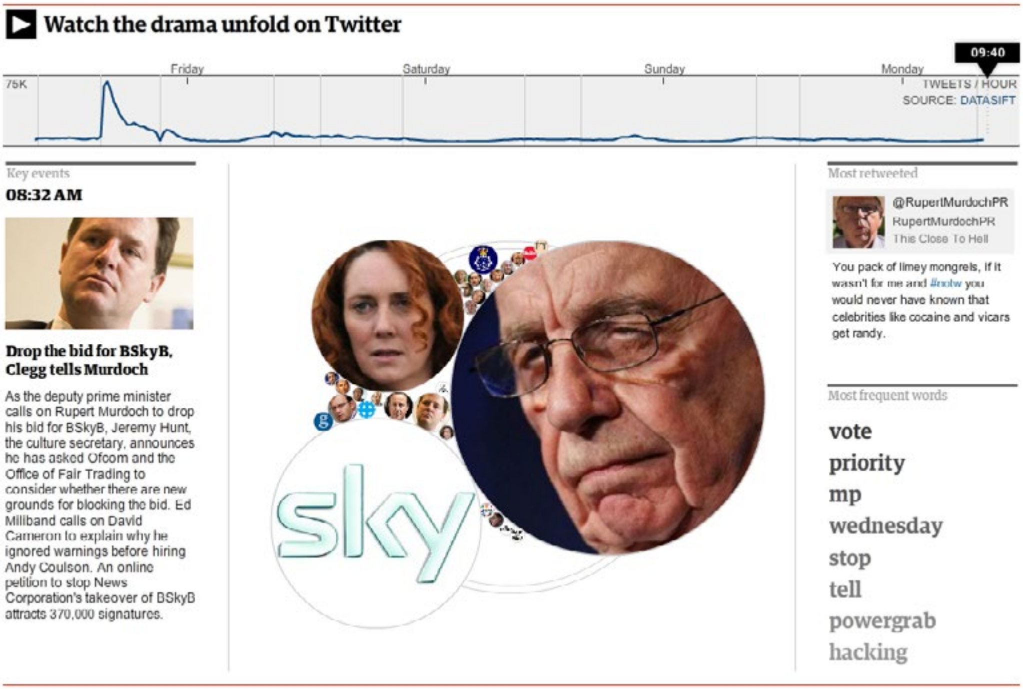
neither can historical tweets be retrieved.[15] APIs thus allow platform owners to structure the possibilities for derivatives of the platform, its use, as well as the access to the platform data. The limits of APIs are increasingly being critiqued, addressing among others their instability, limited access, proprietary APIs and limited data returned (Puschmann and Burgess 2014; Bucher 2013; Gillespie 2010). Additionally, if not everything is accessible via the API we may thus need to resort to screen scraping again.

APIs are thus not simple intermediaries for information but instead mediating objects functioning as 'conduits for governance' (Bucher 2013). The API is arguably a more complex socio-technical device than the scraper as an additional layer of interests and practices, such as those of the third party developer, is involved. The researcher thus enters these complex relations of interests and purposes when using APIs as a source for data. The provision and use of APIs are currently being deployed and developed in a variety of practices and sectors. Compared to industry providing (limited) APIs, scrapers may be viewed as the less polite variant of data collection and in some cases may works against copyright, terms of service, and 'trespass to chattels' (Watters 2011). Despite the significant differences between screen scrapers and web APIs, this chapter stresses how both collection techniques may take analytic advantage of the pre-ordered nature of digital data to both the scraper- and API-assisted researcher.

Scraping and calling APIs, as mentioned before, are not only techniques, but equally involve a particular way of dealing with information and knowledge: they are also analytic practices. In journalism, marketing, and policy research scraping and calling APIs are commonly used nowadays, and much of this activity concentrates on digital platforms offering realtime data, such as the micro-blogging platform Twitter. Scrapers and APIs presuppose a wider socio-technical infrastructure, and the increased availability on the web of 'streams' and 'windows' specifically meant for developers and scripts are especially important in this regard. In other words, the rising popularity of especially APIs is closely connected with the rise of the 'realtime web,' which has been defined in terms of the equipment of web services and digital platforms for the provision of a continuous flow of fresh data (Berry 2011c).[16] Mainly due to its very freshness, this data needs constant disclosure and analysis. Especially APIs are able to capture these fresh digital data and a variety of digital tools and projects, made with APIs, offer realtime analysis. Figure 2 provides an example of the use of an API in 'live'

[15] Through GNIP, one of the largest social data resellers bought by Twitter in 2014, historical and potentially unlimited data are for sale (GNIP n.d.).

[16] The realtime web is closely related to its technical developments (for example AJAX, feeds, push notifications, APIs), and is most prominent in Twitter, Facebook's news feed and Google's increasing preference of freshness in search results. See Chapter 3 for a more detailed history and construction of realtime.



Drop the bid for BSkyB, Clegg tells Murdoch

As the deputy prime minister calls on Rupert Murdoch to drop his bid for BSkyB, Jeremy Hunt, the culture secretary, announces he has asked Ofcom and the Office of Fair Trading to consider whether there are new grounds for blocking the bid. Ed Miliband calls on David Cameron to explain why he ignored warnings before hiring Andy Coulson. An online petition to stop News Corporation's takeover of BSkyB attracts 370,000 signatures.

You pack of limey mongrels, if it wasn't for me and #notw you would never have known that celebrities like cocaine and vicars get randy.

Figure 2: How Twitter tracked the News of the World scandal. Interactive visualization showing the amount of tweets per minute with the #notw hashtag. Source: *Richards et al. 2011*.

news reporting: it is a visualization of the 'phone hacking scandal' from the Guardian Data Journalism Blog, based on the analysis of Twitter data from July 2011. The dynamic version of this visual shows the talking bubble-heads expanding and shrinking in size according to the frequency of their being mentioned on the digital chatter channel, thus providing an account of the scandal 'as-it-happened.'

Morph.io[17] is another example of the type of analytic practice enabled by scraping: it is a platform for developing and sharing scrapers, from which the example in Figure 1 was taken. Morph.io turns scraping into a service, as the platform is built both for programmers wanting to make scrapers and for non-programmers looking for specific types of fresh digital data, such as research journalists. On its site Morph.io explicitly frames its services as including data provision for media, organizations and government. Morph.io also responds to a common problem with scrapers as well

[17] In a previous version of this chapter we used ScraperWiki available at (ScraperWiki n.d.). The public, free, version of ScraperWiki has been discontinued. The example was thus ported to Morph.io, which is very similar to ScraperWiki classic.

as scripts calling APIs: their instability.[18] Scrapers are often custom-built, designed to extract specific types of data from the web, and may also need to be adapted in response to changing access settings or layout and design of the pages to be scraped. This is why many scripts lead a somewhat ephemeral existence, as they are taken in and out of use depending on arising needs, which has been referred to as occasioning the rise of 'live methods' (Lury 2012). In an ostensible effort to address this situation, Morph.io provides a web-based, generally available platform that makes it easier to use, develop, archive and manage scrapers. This also suggests that it would be wrong to approach scrapers as stable, stand-alone machines: scrapers come in and fall out of use; they work, and then they no longer work.[19]

Scrapers and APIs, then, are a multi-faceted phenomenon. They offer techniques to automatically capture digital data, but they are also important components of a broader analytic practice, which is marked by the rise of the realtime web, and the dynamic kinds of information service they enable. Scrapers and APIs seem to imply distinctive approaches to knowledge-making. They arguably come with an epistemology built in: scrapers and APIs format the process of data collection and analysis as an operation of extraction from the negotiated interests and cultures of use. One could say that the combination of these different features makes scrapers and APIs not simply technologies, but socio-technical devices. The various features of scrapers and APIs are of interest in and of themselves, but they are especially important to understand how they can be deployed in digital research—empirically, analytically and normatively speaking.

Scrapers and APIs in digital research

It is also clear that scrapers and APIs are not native to social research. When using scrapers and APIs in digital research techniques from the worlds of information science and digital media are imported; these must be adapted to suit the purposes of social research (just as scrapers and APIs themselves are devices for repurposing

[18] As Helmond (2010) notes, 'Internet methods are incessantly volatile due to the update culture of the Internet itself.' See also Chapter 5 on the volatility of Google as a research engine.

[19] APIs, and especially versioning of APIs, have this problem too. For example, the Netvizz research tool that makes use of the Facebook API and was developed by Bernhard Rieder (2013), had to move from Facebook API 1 to API 2 in 2015 (Rieder 2015). Some features are added, but other important ones for research (such as getting the personal information from everybody in a group) are phased out. For an overview of Facebook's API changes see for example (Facebook n.d.). Look for 'deprecated' to see what was phased out when.

digital data). In addition to the industry roots of APIs mentioned above, the precarity of using scrapers as techniques of digital social research is underlined by the fact that they are principally known in some circles as a technique for e-commerce. There are 'spam sites' or 'scraper sites' that use scrapers to duplicate or recycle digital content—of which Google among others is very critical—and these sites, as their name indicates, commonly use scrapers (Naumann 2015). In some digital discourses, moreover, scraping is most closely associated with illegal practices. For example, the Wikipedia article on scraping is categorized under 'spamming', and scraping is also discussed in these terms on Google blogs; one of Google's larger algorithm updates, Panda, was specifically designed to combat aforementioned scraper sites (see Chapter 5 for more on the relationship between Google algorithm updates and result manipulation). Digital researchers are therefore currently seeking to adapt scraping tools and services to their own purposes as one set of applications of scraping among many others. When taking up scrapers, digital researchers find themselves in the position to adapt a technique to their needs that already serves a broad range of purposes in social life, many of which are not exactly reputable.

One example is info.extractor, a script developed by Chirag Shah and others for the extraction of public comments from the famous digital social media platform Facebook for discourse analysis (Shah and File 2011). While the methodological purpose of this script is academic, the way it works resembles the techniques developed under the category of 'Facebook analytics' by software developers, whose interests often do not qualify as academic (but rather resemble those associated with the Morph.io discussed above). Arguably, the more interesting uses of scripts in digital social research do not disavow these continuities between scripts as a social and computational practice and their applications in digital research, but to the contrary, seek to take advantage of them, in an analytical and empirical way. Some digital research applications of scripts are very explicit about the fact that they are repurposing popular technology. A notable example here is the Google Scraper, a tool for digital textual analysis.[20] This scraper pulls information from the Google search engine, extracting parts of its query return pages.

The Google Scraper has been explicitly presented as a way of repurposing Google as a 'research tool' (see Chapter 5; Rogers 2009). This scraper provides a way to collect and analyze Google return pages, as it allows researchers to automatically query Google for specific keywords and save the results. As such, the Google Scraper and allied tools offer means to repurpose the search engine. Figure 3 provides a snapshot of an initial output of a Google Scraper, which shows the presence and occurrence frequency of issue terms on a particular web page, in this case, that of an advocacy

[20] Google Scraper is, less transparently, also known as the Lippmannian Device.

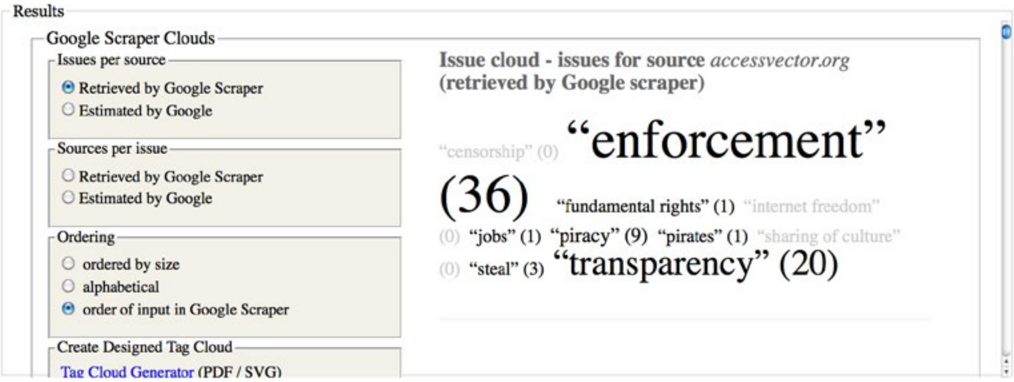
organization campaigning against a European Intellectual Property Law, called ACTA, in July 2012.[21] In this application, the Google Scraper explicitly repurposes available technologies and data for digital research. As such, it also brings into view some of the wider methodological issues scraping poses especially for digital social research. The Google scraper shows that scraping raises potentially very tricky issues for digital research. The question arises to what extent we are studying the medium being scraped, in this case Google, or the forms of social life supposedly generated by this medium. Scrapers raise the question how we establish the difference between researching the medium and researching the social. How far should we go in taking digital devices into account as notable components in our research? Can we understand them as part of our 'methodology', or should we recognize that they are part of our analysis' 'object'?

Scrapers and APIs and medium dependency

The Google Scraper highlights a particular conundrum that since long occupies social researchers: from where does social research derive its analytic categories? From social theory or from the social practices under study? In the past, social researchers had a strong position on this issue, claiming that researchers should not assume that the social world fits their own categories, but should 'follow the actors', and carve up the world as they do (Latour 2005). Arguably, digital research displaces this debate about the provenance of the analytic categories onto the sphere of devices: most notably by 'following the medium' (Rogers 2009), the question becomes how the categories and formats, implicit in digital technologies, structure digital data and analysis. The scrapers and APIs discussed above derive at least some of their formats from the digital media and content queried by them: the table with towns and locations on Wikipedia, the comments on Facebook, tweets, or the ranked page lists of Google. Scraper- and API-enabled digital research tends to adopt analytic categories that have acquired saliency both in the technologies it deploys and the practices these technologies enable.

Scrapers and APIs underline the question to what extent we are studying the platform or engine being targeted or the forms of digital culture and social life supposedly enabled by these digital media (see also Chapter 5 and 6 on this issue). Scrapers and APIs thus draw attention to the question how we establish the difference between researching the medium and researching the social. Scraping and APIs suggest that this

[21] This output is a small part of a larger scrape, which used Google to query 59 pages for the issue terms list related to 'ACTA', presented in Figure 3.



Results

Google Scraper Clouds

Issues per source
● Retrieved by Google Scraper
○ Estimated by Google

Sources per issue
○ Retrieved by Google Scraper
○ Estimated by Google

Ordering
○ ordered by size
○ alphabetical
● order of input in Google Scraper

Create Designed Tag Cloud
Tag Cloud Generator (PDF / SVG)

Issue cloud - issues for source *accessvector.org* (retrieved by Google scraper)

"censorship" (0) "enforcement" (36) "fundamental rights" (1) "internet freedom" (0) "jobs" (1) "piracy" (9) "pirates" (1) "sharing of culture" (0) "steal" (3) "transparency" (20)

Figure 3: Google Scraper result displaying the number of web pages mentioning ACTA issues on the website *accessvector.org*. The issues 'enforcement' and 'transparency' are found the most on pages of accessvector.org. Screenshot taken from *the Google Scraper tool* on February 2, 2012. Source: Google Scraper 2007.

distinction is much less stable or robust than we are inclined to think. As techniques of digital research, however, scraping and calling APIs have some distinctive affordances that seem to make these complications well-worth the trouble. First and foremost, scrapers and APIs solve a problem that digital research shares with many other digital practices: it offers a solution to the circumstance that data on web pages and platforms is not offered in a format that is easily usable. This is why scrapers and APIs allegedly do no less than unlock the 'sociological potential' of the web: they promise to make available for digital research the very large quantities of user-generated data currently being amassed through digital platforms and engines. Crucially, however, is the fact that the popularity of data collection is affecting these very opportunities, as more and more digital media institute ways of regulating access to their otherwise 'generally accessible' platforms, mainly by offering APIs. In this way, the chatter platform Twitter currently regulates access to its data through the Twitter-provided APIs. At the time of writing the Twitter REST API, for example, has a 'theoretical maximum of 3,200 statuses' and this may be subject to change at the company's will (Twitter Developers 2015). While social media platforms promise to make available a wealth of user-generated content or 'social data,' the way they place constraints on access to these data may end up placing constraints on digital social research. Digital research that relies on APIs risks rendering itself platform-dependent, and in effect accepts the black-boxing of its data collection methods (boyd and Crawford 2012).

Scraping and calling APIs are also attractive for digital research for another reason: it may potentially solve the long-held research problem raised by digital data, often referred to as the problem of 'dirty' data (Bollier 2010; Rogers 2013b). Web data collection is often discussed in terms of the onerous and difficult task of having to process ‘incomplete’, ‘messy’, and ‘tainted’ data (Savage and Burrows 2007; Uprichard 2013; Thelwall, Prabowo, and Fairclough 2006). Such statements make sense if we compare digital data to the data sets that many researchers are most used to working with, such as social survey data and interview transcriptions. However, from the standpoint of scrapers and APIs, these characterizations of digital data are decidedly odd. As noted above, scrapers and APIs offer ways to extract structured information from heterogeneously formatted digital data. From this standpoint, it seems wrong to state that ‘digital data’ is inherently this or that. Scrapers and APIs highlight that the quality of digital data is in part a processual accomplishment: it partly depends on the operations devices and researchers perform on web data, how good or clear they are. In this respect, scrapers and APIs also re-open the debate about the techniques used in digital research for protecting and/or enhancing the quality of data (Webber 2009; Gross 2011). The quality of the date may well be considered contextual and dependent on the research objective. Moreover, recent contributions state that scrapers and APIs may ‘“operationalize” the messiness’ of data as something more than accident or inconvenience and in so doing open up new lines of inquiry (Liu, 2015; also see Moretti, 2013).

Scrapers and APIs invite to re-frame or re-locate the researchers’ concern about the quality of digital data. The widespread application of scrapers and APIs across digital culture and social life reminds us that the problem of ‘dirty data’ is not at all exclusive to digital research. Many professions are finding solutions for this problem. This does not only mean that digital research is likely to be scooped in its efforts to make digital data amenable to research (Rogers 2009). We may rather have to define the very field of digital data creation, management, disclosure and analysis in terms of on-going processes of data formatting and extraction. Digital research shares the challenge of how to extract tractable data with heterogeneous sets of other digital actors and agencies (Marres 2012). As the Google founders explain in their classic article ‘The PageRank Citation Ranking: Bringing Order to the Web’: the web is a vast collection of ‘completely uncontrolled heterogeneous documents’ both in terms of internal variations and external meta information (Page et al. 1999). It was in this context that scrapers (and crawlers) emerged as devices capable of bringing order to—or, rather, extracting order from—the web, as they make it possible to collect and restructure large quantities of heterogeneous sources to be queried.

This is what makes scrapers and APIs into ‘devices’ of digital research. If we approach digital research from the standpoint of digital data extraction, it becomes clear that academic digital research shares both its devices and its research challenges with a host of other actors, technologies and agencies that constitute the digital networked environment. Scrapers and APIs are indicative of a wider ‘re-distribution’ of social research that is enabled by digital technology (Marres 2012): many entities and issues that are conventionally located outside social research contribute actively to the performance of digital research. This standpoint has implications for how we understand the relations between the inside and outside of digital research. First, it suggests a relatively broad definition of digital research. Rather than considering digital research as a practice strictly demarcated from other forms of digital data processing, scraping and APIs invite us to approach digital research as a relatively open-ended practice, involving the deployment of a range of digital media devices and practices for structuring social and cultural data, in themselves not necessarily unique to social research.[22] Entities in some respects alien to the context of academic social research may come to play a noticeable role in its organization, with implications for analysis.

Scrapers and APIs also suggest a particular take on the relation between the objects and methods of digital research, and this is where the concept of device proves especially useful to highlight the relatively fluid distinction between object and method. Once we approach digital research from the standpoint of its apparatus, the distinction between research techniques, methods, and objects becomes hard to sustain. In these devices object and method are entangled beyond the point of repair. Scraping and calling APIs provide fitting illustrations for this argument (see also Rieder and Röhle 2012). In the case of scraping it seems impossible to make a straightforward distinction between the instruments, methods and objects of digital research. The Google Scraper, for example, outputs its data as a structured text file and a resonance or word frequency analysis. In this respect, it seems technically wrong to call this scraper a data extraction device: it also analyses and contributes towards the public presentation of results (in the form of tag clouds) (see Figure 3). More generally speaking, it seems strange to ask whether terms identified by scrapers exist 'independently' from the devices deployed. However, scraping not only offers a useful demonstration of the entanglement of methods, object and technique in digital research. It also provides opportunities to deploy them in research.

It may be stated that when scraping and calling APIs involve the importation of digital devices, data and data-formats into digital research, it may enhance the analytic capacities of digital social research. As mentioned, scrapers and APIs tend to derive their analytic categories from the digital platforms and engines queried by them. Scrapers can be used to format and structure data that are implicit in digital data.

[22] The same can surely be said of other social research methods. For example, classification practices of surveys or censuses draw from multiple governmental and commercial communities of practice (see for example (Bowker and Star 2000). It is crucial here that scraping and calling APIs are noticeably open-endedness practices, and that this has consequences for our understanding and deployment of digital social research methods.

This state of affairs, it seems, may be deliberately deployed in digital research for analytic purposes.

Device-driven research and pre-ordered data

In order to clarify what is distinctive about device-driven digital research, it may be helpful to distinguish this approach from prominent other definitions of digital social research. Indeed, as 'medium-specific' digital data capturing techniques, scraping and calling APIs feature prominently, though very often implicitly, in current programs and debates about digital research. In recent years, the key term proposed to capture the opportunities and challenges of digital research as 'big data' was meant to capture the newness and distinctiveness of the mode of research enabled by digital data (Bollier 2010; Manovich 2012b; boyd and Crawford 2012). Although not always specifically mentioned in programmatic writings associated with this term, scrapers and APIs seem nevertheless indispensable. Scrapers and APIs are by no means the only, albeit especially prominent techniques through which big data is delivered, and they have been key to the development of digitized and digital methods. However, big data also indicates what is distinctive or innovative about scraper- and API-enabled social research. After reviewing big data, it will be positioned in relation to device-driven digital research, which suggests to purposefully deploy a particular aspect of scrapers and APIs in social research: the fact that it offers pre-ordered data for digital research.

The research program receiving much popular attention focuses on very large data sets and seeks to push social research to the largest possible scale.[23] This research is often characterized as 'data-driven' (see for example Manovich 2012b; Lazer et al. 2009) as it emphasizes the size and freshness of data sets available for analysis using digital tools of data capture. It also seems mainly oriented towards exploration as big data research tends to 'dig' large data-sets sets for various 'patterns', but the precise nature of the individual patterns recognized may, at least in some cases, seem less important when compared to a more general demonstration of the potential analytic capacity of this type of research. Thus, the study with the title 'One million manga pages' (Yamaoka et al. 2011; Manovich, Douglass, and Huber 2011), and other cultural

[23] Big data refers technically to data that is too big to be processed by current tools and methods (Manovich 2012b), or phrased differently, ;when the size of the data itself becomes part of the problem' (Loukides 2010). It has also been argued that it is not the size of big data that is most notable, but its relationality to other data: big data is fundamentally networked (boyd and Crawford 2012).

analytics projects of the US-based software studies group, was mainly characterized, both in the project publications and by its commentators, in terms of the size of the data-set and style of 'pattern recognition' research.

Device-driven digital research elaborates on, and in some respects may be opposed to, the big data research agenda, insofar as it valorises not so much digital data as the analytic affordances of digital media (Savage, Ruppert, and Law 2010; Rogers 2013b). It is indeed the increased importance of digital media for data analysis that opens up opportunities for social research: they are said to enable distinctive modes of analysis that are indigenous to the medium. Take for instance the method of cross-spherical analysis, which relies on the web to render heterogeneous organizational and media networks amenable to comparative analysis.[24] Cross-spherical analysis seeks analytical cohesion and at the same time respects device-specificity by taking into account how information and issues are structured differently by different digital engines and platforms. In so doing, device-driven digital research arguably opens up an alternative to big data research, insofar as it highlights the opportunities digital media offer for deriving significant findings from relatively *small* data sets and offering *thick* methods by including the operations of digital media into analysis (see Chapter 1). Digital research, in other words, is to a large extent driven by research design (formulating researchable questions, delineating of source sets, developing a narrative and findings).[25]

'Realtime research' enriches the debate about digital research, and points towards specific modes of research enabled by scrapers and APIs. This term has been proposed by Elmer, Back, Lury and Zimmer to characterize the transformation of space and time in social research in the context of digital culture: the increasing valorisation of instantaneity and liveness, the drive towards the condensation of past, present and future in social research, or the evocation of an 'eternal now' (see on this point also Uprichard 2012). But this chapter suggests that 'realtime research' is all about how the digital refigures empirical research. The digitization of social and cultural life, and scraping and calling APIs more in particular, marks that not just time and space in social research, but also *the empirical cycle* itself is reordered. As noted above, scrapers and APIs tend to deliver pre-ordered data, and seem very much entangled in the process of data collection and analysis. Scrapers and API scripts do not just collect data but analyse them in one and the same go, as they parse the data culled from the

[24] See for example comparing the resonance of Issue Animals in the news, web and blogosphere see Figures 25-28 in Chapter 6.

[25] Which is also to say that the delineation of source sets here figures as a key feature of research design, going against the big data idea (or fantasy) of working with 'entire data-sets' and to have done with demarcation, or indeed the problem of generalization.

web. And as already noted, to extract data the scripts rely on information structures embedded in the media, such as the hashtags identifying twitter posts. It might therefore be said that in scraper- and API-enabled forms of research, the identification of relevant pieces of interest for research design tends to precede data collection, rather than succeeding it.

The crucial question is how the pre-ordered data sets are dealt with. Is the content cleaned or stripped of its formatting before analysis, or are these formats treated as *thick* data and thus treated as instrumental in operationalizing analysis? The latter seems to go against a particular assumption in debates about digital data, for example that digital data sets must be characterized as unordered. This assumption seems to be an artefact of the vantage-point from which researchers tend to approach digital research, namely as a form of research that is to be compared to experimentally controlled research such as sample-based survey research and the analysis of textual corpora (Savage and Burrows 2007; Newman, Barabási, and Watts 2006). Compared to complete databases of single source survey or interview data, digital data sets may seem unordered, unruly, incomplete, and unsystematic. From the standpoint of the digital data deluge, however, the opposite seems true: because of their size and freshness, digital research must rely on highly specific markers (or 'ordering devices') present in the data like links, rankings, date stamps, and hashtags in order to keep track on data (see on this point also Langlois and Elmer 2013). Hence the device-driven perspective defines digital research, as research that seeks to derive its analytic capacities from the pre-formatting that is distinctive of digital social and cultural data.

Pre-formatted data as the new social data?

Rather than pursuing these implications on a general and abstract level, however, I flag a specific methodological consequence of doing digital research with the type of formatted data that scrapers and APIs help to extract. Scrapers and APIs enable a style of research in which data features like their date-stamp, their location, and other source attributes such as freshness or connectedness, provide key indicators. I coin these data attributes 'research affordances' which encompass the conditions of possibility advanced in software studies in a wider digital research context.

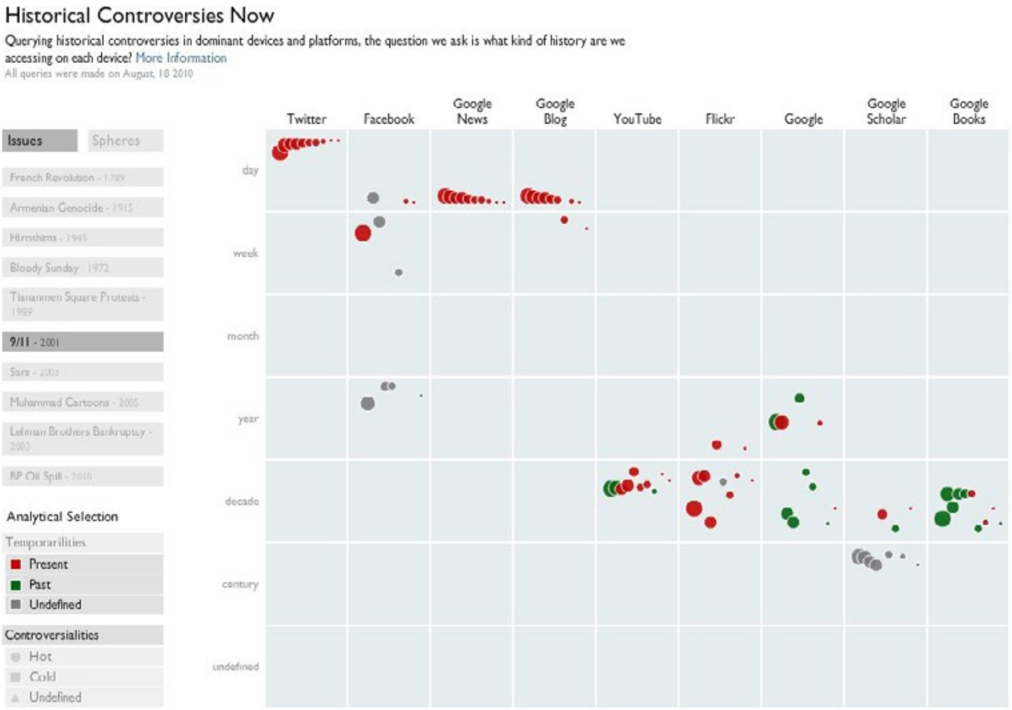
In such research digital media *pre-formatting* data for social research have an operative value. As seen before, search engines format heterogeneous data sets by collecting and processing data through digital markers such as hyperlinks and anchor texts, and digital platforms pre-format social transactions in software and interface design. In both cases the digital media already order, or more precisely, pre-format, the data for digital research. Instead of treating these formatting effects as a contamination of data with noise or negative bias, digital research might derive its analytic capacities in part from these effects. Data formats already implicit in the data may be analytically deployed to structure the research and generate findings. Indeed, in as far as digital data are already pre-formatted by these ordering devices, device-driven research is done with big data: digital media serve as devices to leverage large data.

An example of this approach can be found in the project Historical Controversies Now, which compares ten more or less recent historical controversies across different platforms—from Twitter to Google Scholar (see Figure 4).[26] The project claims that digital media collapse the present and the past into an algorithmic 'nowness,' by examining the ways in which digital media deal with time when structuring information about key historical events. Querying the search engines of several prominent digital media such as Twitter, Facebook and Google Scholar, about selected controversies (for example [Tiananmen Square], [9/11]) the study identified different temporal envelopes for the different media.[27] The micro-blogging platform Twitter, for example, organizes tweets by freshness (i.e. date stamps), thereby favoring re-workings of past events. By contrast, Google Scholar highlights well-cited accounts of the same historic events, which significantly expands the temporal frame, as citations take time to accrue. The study demonstrates how digital media format issues using data points (date-stamps and citations) and how the life cycle of data itself becomes an object and vehicle of analysis. The temporal envelopes of issues across digital media are identified by plotting the publication dates of sources associated with engine returns for each of these digital media. This is an example of medium research, which pursues 'medium findings' instead of 'social findings' as is the case in social research. The difference in perspective is unfolded in this and the following chapter.

Seeking to take advantage of, and aligning itself with, data formats that are inherent to the medium, it is important to note that digital research offers a particular way of dealing with the epistemic issues raised by scrapers, APIs and digital research in general. As we have seen, many of the characteristics attributed to digital data are widely perceived to problematize its analytic affordances and epistemic status. Indeed, critics of big data and digital methods have latched on to precisely these features, in an attempt to challenge the implied naïve empiricism. These issues include but are not limited to the previously discussed instability of data sets, medium dependency (data-

[26] The project is a product of the DMI Summer School (2010) in a project week on web temporalities led by the author. Research in collaboration with Demet Dagdelen, Martin Feuz, Marije Rooze and Thomas Poell. Visualization by Marije Rooze.

[27] Following the dominant query medium, I use [and] to mark the beginning and end of queries (Cutts 2005).



Historical Controversies Now

Querying historical controversies in dominant devices and platforms, the question we ask is what kind of history are we accessing on each device? More Information

All queries were made on August, 18 2010

Figure 4: Historical Controversies Now. Screenshot from interactive visualization showing results for the query [9/11] from Twitter, Facebook, Google News, Google Blog, YouTube, Flickr, Google Web, Google Scholar and Google Books, 18 August 2010. The y-axis shows from top to bottom: day, week, month, year, decade, century, undefined. The X-axis shows results for particular platforms. The markers are either circles for hot controversies, squares for cold controversies, and triangles for undefined types of controversies. The color of the marker is red if it is a present controversy, green if it is past controversy, and grey if the controversy is undefined. The figure shows that Twitter displays recent content about 9/11, while for example Google books displays older content. For the full interactive version see *Dagdelen et al. 2010*.

centrism), and the black-boxing of method. Danah boyd and Kate Crawford (2012) have indeed drawn attention to the relative obscurity and disappointing character in general of the 'data hoses' offered by platforms like Twitter and Facebook, noting that what is presented as a big hose offers little more than a sprinkle; their remarks also amounts to a nice commentary on the inbuilt gender bias in these debates. Others have, in an equally pleasurable way, highlighted problems of data-centrism, as in the comment that the digital social researcher is like a drunk who looks for his keys under the lamppost because that is where the light is brightest (Slee 2011). I suggest that one of the advantages of the device-driven digital research approach is that it entails a change in status of this type of epistemic trouble.

With a view to digital research focusing on research affordances, the structure and dynamism of digital data must be regarded in generative terms, rather than in the negative terms of a bias preventing us from seeing clearly. Data is amenable to analysis precisely to the extent that it is ranked, linked, stamped and tagged, and dynamically so. More generally speaking the epistemic trouble—medium dependency, instability of data sets, black-boxedness—does not necessarily discredit digital research. These issues do not threaten digital research 'as if from the outside', but trouble it in a much more immanent fashion, and may also become an object of research. By plotting the distribution of returns, the above case study Historical Controversies Now, for instance, turned medium dependency—the ways in which digital platforms structure the space-times of inquiry—into a topic of investigation. The epistemic issues associated with digital research mentioned above are thus relevant well beyond the confines of social research: they closely resemble many of the issues that trouble digital culture and social life much more in general. The volatility of media, data-centrism, medium dependency, the opacity of knowledge technologies, are not just problems of social research, but of the information society at large. The epistemic trouble generated by scrapers and APIs are then not just problems of social research, but rather those of many practices that involve the collection, management, analysis, and operation of digital data. In this context, the 'best' attitude towards these epistemic troubles might not be to try and resolve these issues once and for all, or to make them disappear, but rather, to render these problems researchable, to make their effects visible and reportable for practical purposes. It is at this point that the research aims of medium research and social research fold into each other by rendering the specificities of the medium productive in a digital research context.

The issue of pre-ordered data has of course often been treated as an epistemic, normative and political challenge for social research (Bowker and Star 2000; Gitelman 2013). In the context of the web, however, this circumstance has specific analytic affordances. Some academics criticized realtime as a feature of digital social data, flagging difficulties in terms of the limited life span and deterioration of 'live' datasets (Uprichard 2012). However, scrapers and APIs may turn this analytic vice into a virtue. The digital life cycle of may be deployed in digital research; such 'life signals' can be analyzed with the aid of scraper and API techniques. The dynamism of digital data, far from being only a problem (deterioration, incompleteness), may then itself be deployed for analytical purposes in digital research. Concerns about the dynamism of social data may be much less applicable to digital data than some academics seem to assume instinctively. The problem of dynamic digital data does not necessarily have to be sidelined by freezing a dataset in time, so as to render it stable (Latour 2005). In the remainder of this chapter, this proposal is substantiated by outlining a distinctive type of digital research that specifically feeds on the realtime quality of digital data: realtime research.

Liveness is the term proposed by Celia Lury (2012) and Back and Puwar (2012) to investigate the transformation of space and time in social research in the context of digital culture, capturing the need for social research to become responsive to contemporary changes in the spatial and temporal ordering of social phenomena: the increasing valorization of instantaneity and liveness, the drive towards the condensation of past, present and future in digital networked media, and the evocation of an 'eternal now' in this context. Here it is suggested that digital research may indicate an internal re-ordering of empirical research. The digitization of social and cultural life, and the techniques of scraping and calling APIs more in particular, signals a re-ordering not just of time and space in digital research, but also of the empirical cycle itself.

Case study: from live media to the liveliness of issues

In short, digital research seeks to render the formatted, dynamic character of digital networked data analytically productive. As such, this research practice endorses the dynamism or 'shape-shifting' of digital data, turning it into a research resource and object. A crucial question is then how digital research relates to the research practices associated with the web, as discussed above. As already stated, digital research deploys scraping and API tools in order to capture fresh data about current themes, sources and actors. What makes digital research different? The proposal is that it adapts techniques of digital data extraction to determine not just the 'liveness' of specific terms—how prominent are they in current reporting?—but their liveliness.[28] The key question is not what topics, sources and actors have the most currency at a given moment ('now'). Instead, the crucial question for those researching social dynamics is which entities are the most prominent: which terms, sources, actors are the most active, which fluctuate most interestingly over a certain period (Rogers 2002; Marres 2012)? The chapter is concluded by outlining a case study that further specifies this difference between researching liveness and liveliness.

In order to establish the distinction technically, methodologically and analytically, my colleague Noortje Marres and I collected data from two digital media, Twitter and Google, on a small number of terms—austerity and crisis—that are likely to display both dynamics, liveness and liveliness. As part of the scraping and calling API exercise, data on the relative prominence of these key terms is collected as a matter of course, but the variability of these two key words is carefully established: how ac-

[28] In previous work on issue networks, liveliness was defined as fluctuating actor compositions (over time), which can be read in the presence/absence of hyperlinks (Rogers 2002; Marres and Rogers 2005).

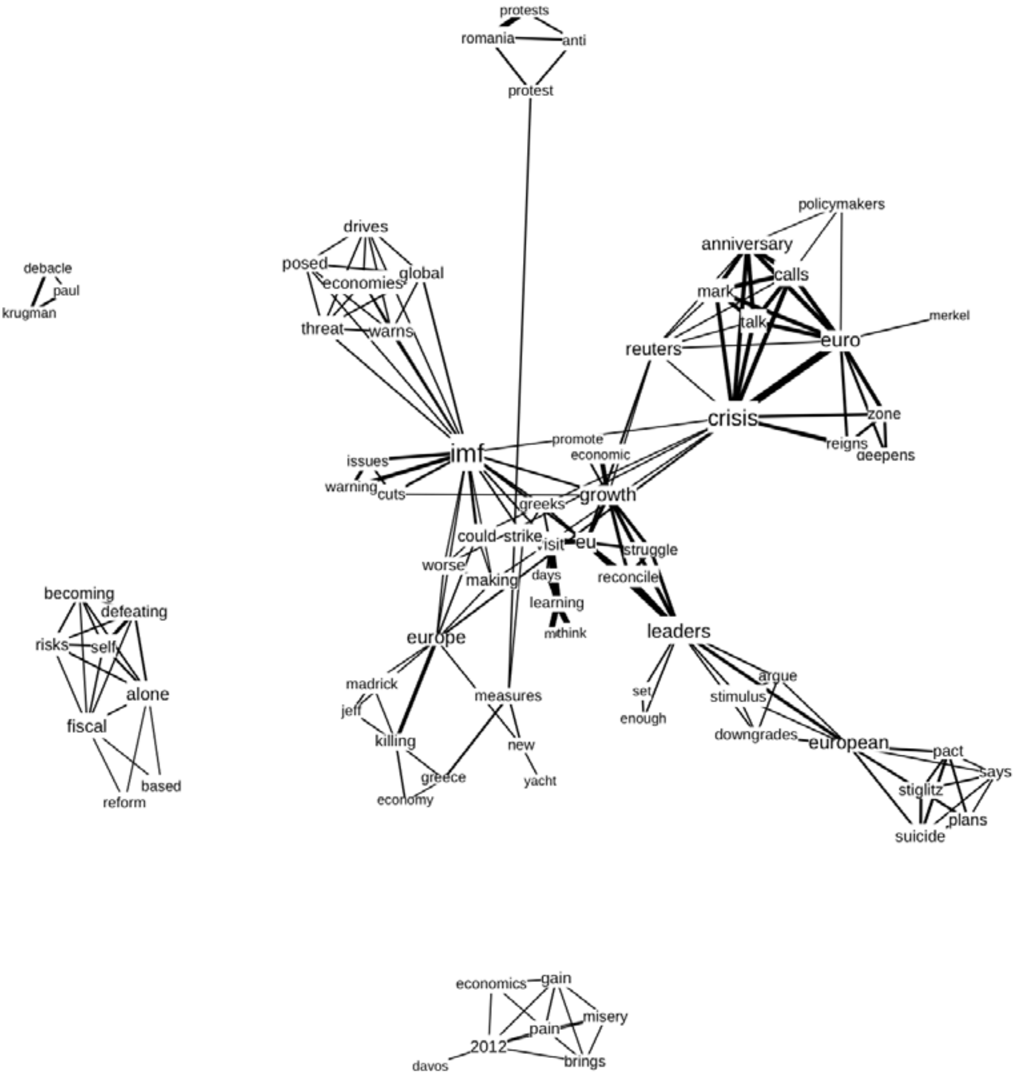


Figure 5: Co-words related to 'austerity' derived from the Twitter streaming API on January 1-31, 2012 concerning over 100.000 tweets. Visualization created by loading the co-word network from DMI-TCAT into Gephi and using a Force Atlas 2 layout algorithm, excluding top 1% most connected terms including austerity and RT, at least 100 co-occurrences are retained, nodes scaled by degree, 2012.

tive or 'lively' are they? Can significant fluctuations in the vocabulary associated with 'austerity' and 'crisis' be identified over time?[29] To make this question operational co-word analysis is used, well-established in both social research and content analysis (Callon et al. 1983; Danowski 1993), and traditionally relying on the formatting of the medium for the structuring of data. Thus, Callon et al. (1983) rely on the keywords used to index academic articles to detect significant word pairings (co-words). In analyzing Twitter and Google data, we also derived the units of analysis from information formats central to the operation of these platforms: the tweet and the hashtag in the case of Twitter; the 'snippet' for Google, a title and string of keywords that Google returns for each individual page in its query return list, and for both platforms the date stamp.

When abandoning to research liveness, and focusing on liveliness instead, a powerful instrument for data reduction is lost. It is crystal clear that asking which term is the freshest, or the most popular, is a splendidly easy way of reducing vast amounts of data to a few significant words. If we want to determine not how topical a term is, but how lively, how do we decide which fluctuations are relevant, in a methodologically sound manner, among the 20.000 word associations we discovered in four days of tweets? Secondly, when exploring the initial co-word maps, by means of the network visualization and analysis software package Gephi, it becomes apparent that the hold of 'liveness' on the respective platforms, Twitter and Google, goes well beyond the prevalence of currency measures in scraper- and API-enabled analyses. The very content of the study reverberated with the 'language of the wire', terms dominating the news of the day, such as greece, imf, debt, bailout, protest, and the cluster '2012 – davos – economics – gain – pain – misery – brings' (see Figure 5 visualizing the 'austerity' issue space on Twitter.)

In an effort to get beyond newsy terms and to more 'lively' ones, we focused the Twitter analysis on co-hashtags instead of co-words, and we reduced the time frame to four days. In so doing we found more lively terms, both regarding the content of the hashtags and in terms of the stronger fluctuations, appearing and disappearing on the co-word maps from interval to interval: #wecanbeheros and #screwyouassad in the case of crisis; and #merkelnotmychancellor; #solidaritywithgreece; #bankstergangsters in the case of austerity (see Figure 6). The assumption that 'social' terms are especially volatile—as compared to terms that figure prominently in the news—is reinforced when Google returns for 'crisis' over time are analyzed. Among the fluctuating terms in this co-word profile are 'planned parenthood' and 'demi moore' and the

[29] Using the Twitter and Google analytics platforms currently under development, an ongoing data collection was scheduled from 1 January 2012 onwards in Google and Twitter for 'austerity' and 'crisis.'

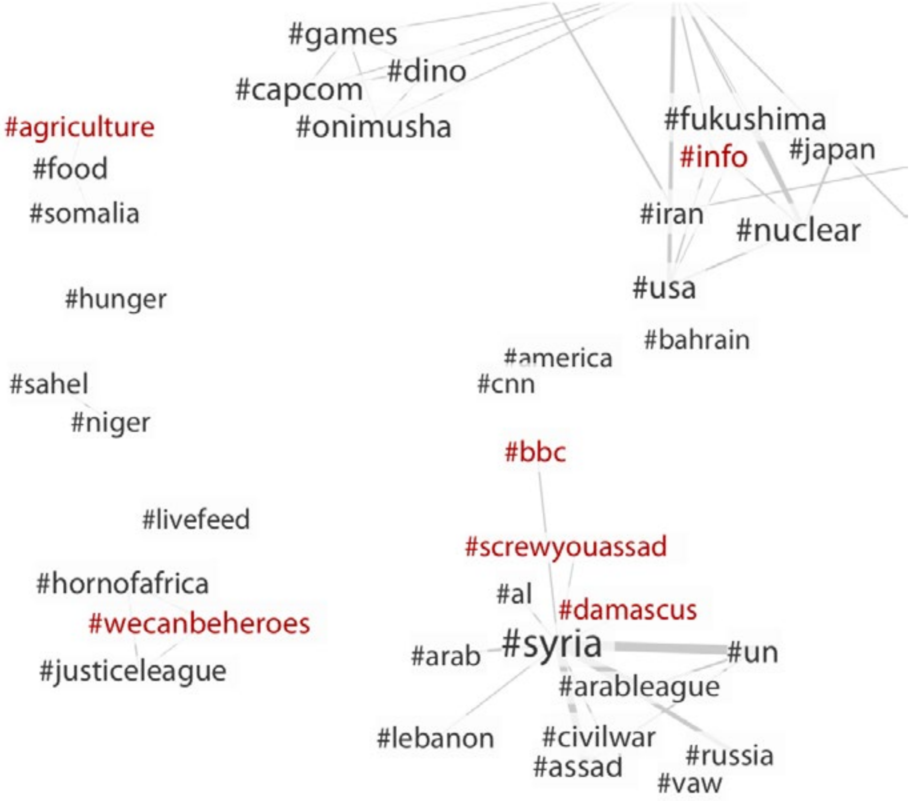


Figure 6: Detailed view of co-hashtag network related to 'crisis' extracted from the Twitter streaming API on January 23-26, 2012 concerning over 200.000 tweets. This visualization shows only static terms, i.e terms that occur across four days in red, and terms that are dynamic i.e. occurring in three days or less in dark grey. The #crisis is removed from the visualization as it connects to all other hashtags. Visualization created by loading the co-hashtag network from DMI-TCAT into Gephi and using a Force Atlas 2 layout algorithm; at least two co-occurrences, nodes scaled by degree, colors were assigned in Adobe Illustrator, 2012.

cluster 'personal; revealing; social; stories' (Figure 7). This contrasts with the more stable political, economic—and more reliably newsworthy—terms that appear across all or most intervals (debt; euro; syria).

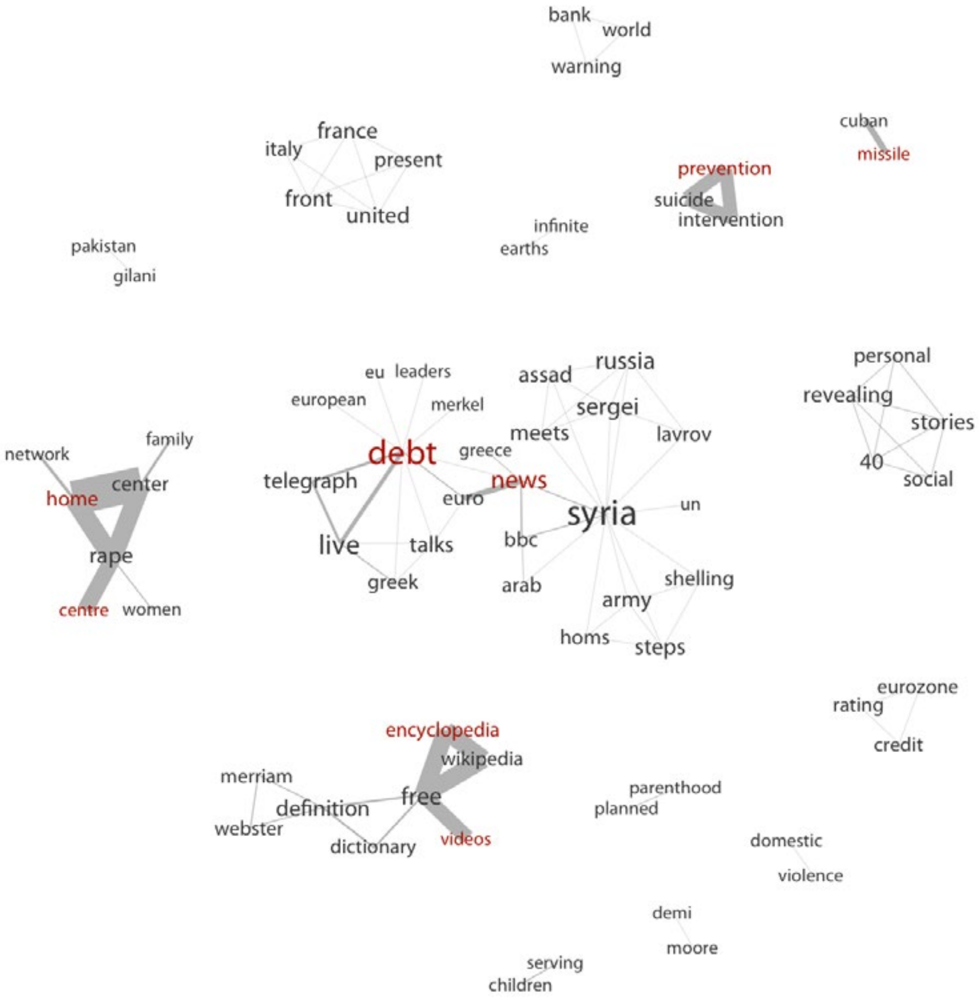


Figure 7: Co-words related to 'crisis' on Google, cumulative view, 1 January - 15 February 2012.This figure shows static terms, i.e terms that occur across all days in red, and terms that are dynamic i.e. occurring in thirteen days or less in grey. Visualization created by loading the co-word network from DMI-TCAT into Gephi and using a Force Atlas 2 layout algorithm; at least two co-occurrences, nodes scaled by degree, colors were assigned in Adobe Illustrator, 2012.

Conclusion

This foray into the temporal dynamics of digital research raises as many questions as it answers, perhaps especially how in digital research digital data can be reduced in analytically meaningful ways. By emphasizing such questions, however, the case study indicates some specific avenues to be explored in the future. No doubt one of the main challenges is how to differentiate more precisely between the study of media dynamics and social dynamics. A distinction between 'scraping the medium' and 'scraping the social' is proposed, which seems to be at the heart of digital research. When the data's medium-specific features are exploited instead of rejected, the question inevitably arises whether social life is studied or rather the media partly enabling it. But the difference between 'scraping the medium' and 'scraping the social' is probably best understood as a difference in degree: in some cases, digital devices play an noticeable role in the structuring of data, while in other cases a discernable empirical object cannot readily be reduced to the medium-architecture enabling it. In scraper- and API-enabled research, the boundary between the media apparatus and the social object must be understood as flexible, moving between 'all apparatus and no object', and 'part apparatus, part object'. This distinction may indeed be regarded as a methodological accomplishment, one of the key merits of digital research.

Lastly, digital research also unsettles the distinction between social research and social life. It particularly challenges the strict separations most of us have learnt to make between the epistemological issues troubling scholars and scientists and the real-world concerns occupying social actors. The methodological and conceptual problems raised by live digital research surprisingly resemble the issues that trouble digital cultures and societies. As discussed, one of the big problems digital research faces is how to gain full and reliable access to data collected on digital media. While not necessarily solving this, digital research requires a change in status of this type of epistemic trouble. In digital research, epistemic issues—medium dependency, volatile media, black-boxedness—trouble research in a rather immanent manner: they affect it from the inside and may become an object of research. The case study Historical Controversies Now turned medium dependency into a topic of investigation. The study of austerity and crisis on Twitter and Google highlighted how data, platforms and research together provoked the preoccupation with liveness. The epistemic issues arising in digital research are therefore not so different from the issues troubling digital culture and social life much more in general.

The politics of realtime

Chapter 3

THE MICRO-BLOGGING platform Twitter 'is a real-time information network' (Twitter n.d.) where 'the real magic [...] lies in absorbing real-time information that matters to you' (Twitter Support n.d.); Facebook's Newsfeed ticker 'shows you the things you can already see on Facebook, but in real time' (Facebook Help n.d.) and search engines such as Bing and Google try to include fresh updates in their result pages so that 'relevance meets the real-time web' (Singhal 2009). This chapter further explores the central position of 'liveness' in digital media, as described in the previous chapter, by proposing ways to empirically study it. Set against the previous chapter, however, this chapter emphasizes medium research instead of social research by using digital media as devices in the method so as to further explore the construction of media streams in order to reach conclusions about the digital media.

The medium research perspective proposed here combines different types of methodological and conceptual resources to study which technologies and data are relevant for digital media to organize realtime, in order to develop a critical understanding of how processing and engagement become operative in digital media. In so doing I connect to one of the core concerns of software studies, namely the specific operations of the digital media in social and cultural life. As an interdisciplinary field, software studies researchers use a variety of materials to study the 'stuff of software' (Fuller 2008, 1), including reading the code (Marino 2006; Galloway 2004, 20), patents (Rieder 2012), legal documents (Van Hoboken 2012), trade press and technical documentation (Bucher 2012a; Bodle 2011; Kirschenbaum 2003), but also by studying settings and interfaces (Stanfill 2015; Bucher 2012a; Gehl 2014). Such varied materials offer different vantage points on the study object and allow to research different aspects of computational objects. For example, interface analysis offers ways to study how digital media produce norms by privileging some uses over others through their affordances (Stanfill 2015). Studying API documentation may provide insights in the interoperability and data circulation in the back-end of platforms (Bucher 2012a; Bodle 2011). Some research in this area uses the digital data of media to investigate the medium, and may be considered as device-driven, but it is a relatively underexplored approach (see for examples Feuz, Fuller, and Stalder 2011; Gerlitz and Helmond 2013; Bucher 2012a; Sandvig et al. 2014). My contribution to this rich area of software studies is what I understand by medium dependency; that is, I focus on how digital media as devices intervene in the method. I do so by inquiring into the configuration of the medium device in the larger research apparatus. The digital methods approach to software studies, which I refer to as medium research, uses the affordances of digital media as the main vehicle to study digital media (see also Chapter 5 on research with Google algorithms where I develop this approach further).

The empirical case study presented in this chapter empirically and conceptually engages with debates around the notion of realtime (Gehl 2011; Berry 2011c; Chun 2011). Realtime has not only emerged in academic discourses on time, but has been especially promoted by industry to label the post-web 2.0 era. The 'realtime web' is therefore the successor of the less well-known 'live web' which sought to break with the 'static web' organized around web pages and links, as initially conceived by Tim Berners Lee (Searls 2005). The term was coined 'to describe the exploding number of live social activities, from tweets to status updates on Facebook to the sharing of news, web links, and videos on myriad other sites' (Hof 2009). In addition, realtime entails the promise of an experience of the now, allowing platforms and other web services to promote the speed and immediacy at which they organize new content and enable user interaction (Gehl 2011). Geert Lovink even goes as far as saying that 'real-time is the new crack' (2012), since engines and platforms increasingly call upon the technological imperative by investing in the optimization of both processing information and the possibilities of user engagement. The notion of realtime is thus used to describe media characterized by fresh, dynamic or continuously processed content as opposed to static or archived media. Non-realtime media feature non-dynamic or historical content and are not designed to continuously process and present the latest or most relevant content.

With this focus on dynamism and change in the present, realtime is also used as a universal temporal container in both academic and commercial discourses, implying a flat understanding of the term. Instead of thinking of digital media as operating in realtime, in this chapter I argue that what is considered realtime is created in specific ways. The objective is to 'unflatten' the nowness of digital media and to empirically study the relationship between engagement and digital media, by introducing a medium research perspective taking into account how various platforms and engines create distinct realtimes. By approaching the notion as a form of information organization the aim is to specify and qualify its assembly. The focus is therefore on the intersection of processing and engagement by tracing the technical features and developments, which historically have been deployed to organize realtime. The role of push and pull technologies are addressed and so are syndication and streaming protocols.

The empirical case study is comparative across a selection of dominant engines and platforms. Each of the devices are configured in such a way that they output the pace, or rate of fresh content, that appears in the streams or result pages. In so doing, the comparative analysis makes visible patterns of updates and the prominence of 'sticky' promotional or other content that is persistently at the top of the stream. In the empirical case study an information-based view of realtime is developed by introducing the notion of 'pace', which is tied to the algorithmic order and presentation of information. In order to empirically explore digital nowness, the actors, infrastructures, features, objects and activities organizing the pace of streams in a selection of platforms and engines are traced. In drawing attention to how realtime is specific and internal to digital media, the focus is not solely on the technology of platforms; the social arrangements and cultural practices they incorporate and enable are also taken into account. Following Niederer & Van Dijck (2010) and Bucher (2012b), the term ‘technicity’ is used to focus on the socio-technical relations producing realtime.

More precisely, in the case study the pace of an issue across search engines and platforms is explored in order to identify how the organization of content, its presentation and user actions produce very specific modes of realtime. The relation between freshness and relevance as modes of organizing realtime is of particular interest, based on the assumption that fresh and relevant organizing principles create different configurations of realtime streams. Media do not operate in realtime. Devices and their cultures operate as pacers of realtime—a perspective which complicates universal accounts of realtime (Berry 2011a; Berry 2011b) and ‘realtime research’ as developed in the previous chapter (Lury 2012; Back and Puwar 2012; Elmer 2013; Back, Lury, and Zimmer 2013). The fabrication of these specific configurations as ‘realtimeness’ is addressed to underline the fact that realtime is not a framework in which media change, but is in itself assembled through the technicity of platforms. In contrast to flat notions of realtime, focusing on ‘realtimeness’ highlights the distributed fabrication of engagement in the present and opens up discussions on the politics of realtime. In so doing, the chapter is an example of studying the medium with the objective to arrive at findings about the media under investigation.

Realtime experience and processing

The notion of realtime within media studies is typically approached from either a user experience perspective or viewed as a computational process, creating a twofold focus on experienced and technical time (Leong et al. 2009, 1277). In his early discussion of digital temporality, Adrian Mackenzie contends:

> real-time concerns the rate at which computational processing takes place in relation to the time of lived audio-visual experience. It entails the progressive elimination of any perceptible delay between the time of machine processing and the time of conscious perception (Mackenzie 1997, 60).

He distinguishes between the realtime processing of information through algorithmic and computational processes and the realtime experience offered to users through web interfaces. Realtime, Mackenzie’s account suggests, is not an external time frame in which events or web engagement occur, but is a fabricated temporal condition, in which the processing of information is organized at such speed that it allows for access without perceptible delay. The real in realtime is therefore not related to immediacy—which is literally impossible—but a question of speed and the organization of content in relation to time. Hence, Mackenzie’s account resonates with computational perspectives on realtime, focusing on operating under time constraints. Here, realtime refers to systems and processes performing tasks in predetermined temporal windows, most notably in micro- or nanoseconds and their computational challenges.

The interplay between experience and processing, Mackenzie continues, is situated at the heart of the Internet and is fundamentally concerned with information production, access and algorithmic sorting, opening up an informational account of realtime. Mackenzie's distinction between realtime processing and experience, however, was developed in the mid 1990s when the web was dominated by mainly static pages and very few platforms and engines. The question therefore is whether this relation needs to be revisited in the contemporary state of the web, which is dominated by platforms, engines and their dynamically updated content.

Paul Virilio also forwards the informational viewpoint on realtime when he determines the value of information based on the speed of access as 'the reality of information is entirely contained in its speed of dissemination [...] speed is information itself!' (1995, 140). Early debates on realtime, situated in the 1990s, are thus fundamentally entangled with the idea of the information superhighway as 'the global movement of weightless bits at the speed of light' (Negroponte 1996, 12), in which the key potential of the Internet was considered to be its information access at high speed. Manuel Castells (2000) examines how speeding up alters the user's experience and suggests that the constant focus on accessing information now limits the focus to the present and isolates it from both future and past. What emerges is a 'timeless time' or a 'non-time' without past and without duration. Although formulated during the early years of the Internet, the critique of 'timeless time' has been revived in recent years in relation to the increasing prominence of streams, which are perceived as encapsulating users into an 'eternal now' (Uprichard 2012). In a similar fashion, 'realtime research' seeking to collect and analyze data 'live' while it is being produced, often faces critical voices suggesting that such attempts of being 'live' may render research atemporal, trapped in the present and merely speed-driven (Back and Puwar 2012).

In relation to the contemporary realtime web and social media platforms, user experience with immediacy and speed is still perceived as relevant, yet in different form. No matter what users do, where and why they do it, there is always a platform inviting them to share that information. Twitter asks, 'What is happening?' and Facebook prompts 'What's on your mind?'; both platforms inform users immediately and in a similar manner when someone else has acted upon their shared activities. Realtime experience is no longer limited to the elimination of a perceptible delay between the request and the processing and presentation of information. Instead it encompasses modes of engagement, interaction and the speed at which responses to one's own actions are shown.

The question therefore emerges, to which extent the social web alters the way realtime experience and processing, as perceived by Mackenzie (1997), inform each other? User front-ends of social media platforms, Gehl argues (2011), are characterized by numerous features for immediate and speeded up content engagement—thus opening up questions about the political economy of realtime media. Whilst user interfaces are focused on immediacy, only platform owners and in part cooperating partners can access and process content in the archival back-end (Stalder 2012). Gehl claims:

> Here, we confront a contradiction: the smooth interfaces that users enjoy appear to be comprised solely of immediate connections and instant information, but the servers powering them are maintained in large part due to their long-term, archival potential. This contradiction is the motor that drives Web 2.0 (2011, 2).

It is the immediate, ongoing user interaction that allows to fill the associated databases with new data, contributing to a constant interplay between '"real-time drives" and the archival impulse' (Gehl 2011, 6); it leads to a dual temporality of web devices to which not all actors have access. Although Gehl's distinction between front- and back-end temporality creates a rather general notion of realtime in front-ends, he underlines how temporality is not external but fundamentally internal to media in a wider sense. However, these operational capacities are not necessarily immanent to the technologies as such, but informed by the politics of platforms (Gillespie 2010), which rely on economically valorizing user interaction and data. Platforms, Gehl's work suggests, are not realtime media, they produce distinct forms of realtime for specific users, which is conceptualized as 'realtimeness' in the course of this chapter.

The most central feature in many discussions around realtime is the stream. Streams are automatically updating content flows and have become key elements of websites and social media platforms (Berry 2011b; Berry 2011a; Borthwick 2009; Lovink 2012; Manovich 2012a): 'A stream is a dynamic flow of information (for example multimodal media content). [It is] instantiated and enabled by code/software and a networked environment' (Berry 2011c). The increasing presence of streams, software studies scholar David Berry (2011c) argues, is closely tied to the rise of the social web where users do not have to search for content on static web pages—as in the 'destination web' (Berry 2011b) or 'static web' (Searls 2005)—but where content is brought to them instantly through automatically updating streams, recommendations and other dynamic elements. Different platforms offer different encounters with stream content, which come at 'varying lengths, modulations, qualities, quantities and granularities' (Berry 2011c, 144).

The various features that organize the experience of immediacy and speed, give rise to a couple of insights. First, the web does not merely change 'in realtime', but actually produces specific temporalities through its engines, platforms and the related web cultures. Second, a clear differentiation between realtime experience and processing becomes problematic, especially in the context of streams, as realtime media content is produced, processed as well as engaged with in realtime, while at the same time the activities of users in the front-end inform the processes in the back-end. Third, the interrelation between processing and experience is increasingly subjected to platform political objectives. In order to account for the specific fabrication of realtime the interrelation between experience and processing needs to be considered which I will do by engaging with a device perspective, exploring the interplay between features, data points, algorithms, practices and uses of digital media.

Contemporary dominant digital media are engines, search engines with their own logic of ranking content or sources, and platforms, each platform hosting and organizing specifically formatted content. A device perspective on these media does not only focus on the technical aspects but considers their operational capacities as informed through the social arrangements, cultural practices and politics incorporated and enabled by digital technologies. As argued in the introductory chapter, platforms and engines are epistemological machines capturing, processing, analyzing, ranking, recommending, formatting and aggregating data on the web. From a device perspective, digital media do not come with clearly delineated boundaries and operations, but inform and are informed by a multiplicity of actors, dynamics and practices. Additionally, as discussed in the introductory chapter, device-driven research emphasizes the device's role as structuring both digital social life and research.

In what follows, I seek to further develop the device approach to realtime research, but as opposed to the previous chapter, I focus on the modes of realtime constructed and produced by different digital media, by differentiating between the modes of engagement enabled in the front-end and their relations to sorting, processing and organizing content in the back-end. When examining realtime as a socio-technical arrangement in the next step, I turn to the technical organization of web realtime, underlining different ways to fabricate realtime and addressing the role of user engagement.

The technicity of web realtime

Although the recent popularity of realtime is closely related to the rise of streams in social media, realtime has a longer history (Chun 2011); it determined how the Internet developed from its early stages onwards. Focusing on the socio-technical relations that compose realtime, this section traces the technical features that allow to automate, speed up, and organize immediate access to new content and provides an indication how realtime is fabricated across both front- and back-end.

In the early days of the web the timeliness of content delivery or information retrieval was often framed in the existing models of ‘push’ and ‘pull’ technologies. In a pull model the web user ‘pulls’ information by querying a website at a chosen time; this transaction is initiated by the user or client himself. In a push model the provider or server initiates the transaction at specific intervals, the content is ‘pushed’ to the user without a specific request (Franklin and Zdonik 1998, 516). In the mid 1990s push technologies were presented as a way to automate requests for new web content, thereby addressing the novel problem of ‘information overload and the inability for users to find the data they need’ (Franklin and Zdonik 1998, 516). This problem arose with the increasing number of web pages that bypassed the browser to retrieve content from the web (K. Kelly and Wolf 1993).

In its heydays (1996-1997) push technology companies such as PointCast and Marimba offered software allowing users to subscribe to specific channels and automatically receive new, pushed content. These updates were delivered in intervals specified by the user or client (Gerwig 1997, 14) to achieve ‘the appearance of having real-time updates’ (Franklin and Zdonik 1998, 516). Such push technologies, however, did not provide a continuous inflow of new content but were scheduled, periodic pulls that simulated the idea of push. While in the pull model the delivery of fresh content was defined by the user, software settings in the push model specified when the retrieval was initiated in predetermined intervals.

The major browsers at the time, Internet Explorer and Netscape, both became actively involved in developing push technologies as standards and software. In 1997 they developed push interfaces to the computer’s desktop, respectively Microsoft Active Desktop and Netscape Netcaster, which no longer required the launch of a browser in order to retrieve updates from the web. To this end Netscape used existing standards while Microsoft developed the Channel Definition Format (CDF) providing a new delivery mechanism for web content that would ‘turn every desktop window into a channel. Instead of a window framing a static page, it frames an ongoing stream’ (K. Kelly and Wolf 1993). However, in the pre-broadband era of the late 1990’s, the push hype ended when these technologies clogged networks by using excessive bandwidth to retrieve updates in intervals (Bicknell 2000). After browser companies Netscape and Microsoft discontinued their push products Netscape continued to develop their underlying standards to syndicate new content for their Netscape Netcenter portal.

With the increasing amount of web pages and the need to navigate between them a number of technologies were introduced to organize and order content on the web such as directories, portals and search engines. These media often use web-native objects such as hits, links or timestamps to organize and rank content. In the mid 1990s the portal became an important entry point to the web as it provided a one-stop destination for relevant and fresh web content aggregated from a variety of sources. Portals marked a shift from an all-purpose static destination web to a more personalized and dynamic web (Steinbrenner 2001, 1). The My Netscape portal is particularly interesting, since it differed from other portals in that it included a newly developed format for publishers to syndicate fresh web content: Rich Site Summary or Really Simple Syndication (RSS).[30] Every webmaster was able to define channels on their website that could be monitored for new content, so that the portal became an aggregation infrastructure for customized and timely content. RSS did not allow to engage with content in continuous realtime, but portals would pull in new content in scheduled intervals by polling blog subscriptions or feeds. Whenever users loaded the portal page, they would receive updated and time stamped content, creating a specific experience of realtime. Users could either manually request or 'pull' new content by reloading the page or automatically retrieve new content by adjusting the refresh rate of the intervals for preferred pages. The portal's new syndication format automatically retrieving new content from external sources presented an early idea of a 'realtime web environment'.

When AOL acquired Netscape in 2001 it stopped supporting RSS. It was then passed on to software developer Dave Winer's company UserLand Software (Winer 2000; Festa 2003) who had experience in developing blog software and had developed an early syndication format called <scriptingNews>. This format, which turns a website into a 'specialized content flow' (Winer 1997), eventually merged with Netscape's similar RSS format. When RSS was integrated in popular blog software such as Moveable Type, it became the default for syndicating blog content. Blogs, displaying the latest content on top with their reverse-chronology, now send out RSS feeds by default and allow to subscribe to new content updates aggregated and read in a feed reader. Besides RSS as an update mechanism, blogs also notify (search) engines of new updates using 'ping'. This mechanism was introduced in 2001 by Winer and is used by (blog) search engines to send a notification to a ping server that can be polled by search engines for new content. These ping servers enable search engines to keep their index fresh with the latest blog posts.

Portals and feed readers use a pull mechanism where the server is periodically polled for new data, but in recent years several technologies such as PubSubHubbub (PuSH)

[30] Initially the format was conceived 'as a metadata format providing a summary of a website' but it became clear that 'providers want more of a syndication format than a metadata format' thereby reshifting the focus on timely content update (Libby, Dan 1999).

and HTTP Streaming were developed to immediately push new content and provide instantaneous updates (Leggetter 2011). Since 2006, with the rise of micro-blogging platforms like Jaiku and Twitter and social aggregators such as FriendFeed immediate updates have been privileged. Returning to the idea of the realtime web, a number of new computation and web server frameworks have been developed, focusing on the realtime processing of information streams including Tornado, developed by FriendFeed, (Recordon 2009) and Storm, used by Twitter (Apache Software Foundation n.d.). With the launch of Twitter and the Facebook News Feed in 2006, streams have become a key element in social media platforms. At the time of writing, Facebook features a multiplicity of streams on its homepage, from the News Feed, which can be further differentiated into streams focusing on 'most recent' content, photos only, close friends only or all friends, to the Ticker which displays even more fine-grained information at high speed, to the possibility to create a series of interaction streams through Facebook chats. Streams, as compared to previous technologies, are technically innovative in that for example Twitter's Streaming API establishes a persistent, two-way connection when a filter (for example hashtag or keyword) is created.[31] Facilitated by Storm, a distributed realtime computation system, Twitter uses stream processing 'to process a stream of new data and update databases in real-time' and continuous computation to 'do a continuous query and stream the results to clients in real-time' (Marz 2011). Whenever a new tweet that matches the filter arrives in their database, a trigger is activated, pushing the tweet into the persistent connection between their server and the user. This mechanism is novel as the previously mentioned ones work on opening and closing connections, polling servers and pulling in information, either on request or scheduled.

Although acknowledging the current dominant account of realtime which focuses on streams within social media, this chapter also pays attention to realtime aspects not covered by the stream, such as the increasing attention for realtime by search engines. Search engines typically do not present their results in a stream, but rather in a ranked list. In the current default Google experience, with Instant the result page updates itself in realtime when users type a query, with realtime suggestions for formulating that query with AutoComplete. Whereas this is a move towards realtime in the front-end, there is also an increased move towards realtime in the back-end of the engine. Google's famous PageRank algorithm measures the relevance of web pages by evaluating hyperlinks, although Google increasingly includes other signals in its algorithm. The various algorithm changes of the search engine are addressed in Chapter 5, but it is important in the context of this chapter that since 2001 Google has introduced algorithm changes in order to become a 'realtime' search engine for

[31] The Twitter Streaming API is based on HTTP Streaming to push updates to a web client by keeping a persistent connection open (Twitter Developers n.d.).

'hot' or 'happening' issues privileging fresh results over relevant ones in terms of their PageRank (Wiggins 2001; Singhal 2009; Singhal 2012). A significant moment in this evolution is the introduction of google.com/real-time in 2009 which enabled the search of social media results, but which no longer exists due to discontinued contracts with the main content provider Twitter (Sullivan 2011). A second significant moment is the Caffeine update to their index, also introduced in 2009. Before the Caffeine update googlebots, Google's software crawlers, were sent out to index changes and new content in scheduled intervals, which meant that for non-news sites Google's main index would be refreshed about once every week. With the new update fresh content in the index is updated almost instantaneously. In addition to the Caffeine update, Google started to increasingly privilege fresh results over authorative results allegedly to comply with the post-September 11 demand for realtime news and trusted websites (Wiggins 2001) by millions of users. Noteworthy is for instance the Query Deserves Freshness algorithm, which is developed to determine whether a topic is 'hot', so that the query may need fresh results on the top of the result page, as in the case of breaking news stories (Singhal in Hansell 2007). Similar to the Google search engine, the Google News index is filled with content that is crawled by googlebot. However, the ranking algorithms of Google News follow the updating cycles of the news sites that are included in the index. The Google Blog Search, which was discontinued in 2014, had an index that included blogs containing an RSS or Atom feed, and were updated by monitoring ping services, including Google's own Blog Search Pinging Service to retrieve the latest updates from blogs including them in their index almost immediately.[32]

Having traced the technologies and standards which inform content organization in the back-end of various devices such as pull, push, RSS, ping, persistent connection and the Google's Query Deserves Freshness update, a device-driven approach to realtime examines these technical standards used by devices, in order to learn how they organize content both algorithmically and by user activities and interaction, how such content can be queried, how it is (re-)presented, and, finally, how they can be repurposed for digital research. This chapter, take up the operative capacities of realtime media to study the construction of realtime and the differences across media. In this chapter, these various features and agents contribute to the pacing of information available on the web, the rhythm in which it is found, retrieved, sorted and displayed both in the front- and back-end. On the basis of understanding the assembly of realtime, the following section addresses how we can empirically qualify the specific realtimes a device fabricates.

[32] More precisely, the Google Blog Search engine was discontinued as a separate search engine and was incorporated into the Google News Search engine (see for example Schwartz, Barry 2014).

Case study: Modes of realtime

The device-driven contribution to realtime is approached by configuring the search engines and platforms in such a way that they output their ‘pace’, or rate, of fresh content appearing in their streams or result pages as a specific mode of realtime. The logic of digital devices to organize information is often related to both the content’s relevance and freshness. Pace, closely related to rhythm or tempo, is a term used to describe the relative speed of progress or change, or the rate of a repeating event; it provides a way to empirically study realtime dynamics. Pace emphasizes the ways in which digital media deliver fresh content. So far, pace has been approached as the rhythm through which a multiplicity of features, agents and practices associated with a device organize, or pace, the flow of new content. In its general understanding, pace has often been linked to pacing devices, in a sports context where pacers or pacesetters control the speed of runners or cyclists during long distance training, or cardiac pacemakers which control the rhythm of heart muscle contraction. Pacing devices thus strategically organize the speed at which movement and change occurs, underlining the collaborative fabrication of speed and time. Beyond a more universal notion of realtime, it can be specified how paces differ; they can be described empirically by focusing on the relation between freshness and relevance for each digital device. It is assumed that freshness and relevance create different paces, and that the pace within each engine and platform is internally different and multiple in itself.

In the empirical study colleagues and I focused on one prominent issue at the time of research in summer 2010: Pakistan Floods.[33] Focusing on an interval of 24 hours (from 15:40 hours (CET) on 18 July until 15:40 hours (CET) on 19 July 2010) we explored the pace at which various devices presented new content through Facebook, Twitter, Wikipedia, Google, Google News, Google Blog Search, YouTube and Flickr.[34] The case study assessed what devices do with the different paces of content. How can these paces be characterized? What is the relation between relevance and pace? And how relevant is freshness? Realtime media are often used as platforms for crisis communication during disaster events (Vieweg et al. 2010; Bruns et al. 2012). This was also the case with the 2010 Pakistan Floods (Murthy and Longwell 2013). This raises

[33] The project is a product of the DMI Summer School (2010) in a project week on web temporalities led by me. The research was conducted together with Anne Helmond, Carolin Gerlitz, Taina Bucher and Erik Borra.

[34] This case study was selected because the Pakistan Floods were a major unfolding news event that we recorded as it unfolded. We captured a day (July 18-19) at a very early stage of the event, which eventually became the #3 News Event in Twitter’s 2010 Year in Review (Twitter 2010). All the selected devices were then dominant (US) web platforms for respectively social media status updates, encyclopedic information, search, news, blogs, videos, and pictures.

questions about the affective nature of the unfolding disaster and ethical concerns about the subjects and suffering involved.[35] This chapter acknowledges these issues but does not directly engage with the affective experience of floods, nor the subjects engaged in or commenting on the crisis as no usernames were published. Instead, the case study is used to research the realtimeness of a current event that unfolds on digital media. For this purpose, we decided to monitor the pace at which new content was provided to users for 24 hours by setting 5-minute intervals.

For each of these devices, queries were designed to fit the platform or engine, since content is organized and offered differently per device. While search engines return results for a *query*, micro-blogging platforms encourage users to organize content around *hashtags*, which are tags prefixed by a hash symbol that can also be used for searching. Engines and platforms have thus preferred entry points or queries.[36] In other words, by querying the digital media we created result lists or content streams for the cross-comparison of pace. We scraped the above-mentioned platforms and engines for the corresponding queries for the 2010 Pakistan floods, taking into account the preferred queries as put forward by the device or its users. For example, Google recommended the query [Pakistan Floods 2010][37] whilst Twitter users decided on [#pakistan] as the most dominant hashtag.[38] Finally, we queried Facebook using its search feature for [pakistan flood] in Posts by Everyone, which returned public mentions of Pakistan flood in All Post Types (Links, Status Updates and Wall Posts, Notes) by everyone.[39] Most digital media offer a (default) result page, which is ordered either

[35] Sara Ahmed's (Ahmed 2004) work on affective economies may be useful in considering how the affective dimension of the flood may be co-produced by specific ways in which realtime media operate, just like constantly updated streams may create new forms of affective proximity. Especially the perpetual, high-paced presentation of new insights, perspectives and information on catastrophic events as offered by realtime devices enables users to experience both the unfolding of the event as well as its media discussion as they happen.

[36] Platforms such as Facebook and Twitter both offer streams of content based on friend or follower lists too. However, to render the engines and platforms comparable we accessed realtime through an event specific query in all devices. See Chapter 7 for a second elaborate exploration of how to make different types of devices comparable.

[37] Google Autocomplete suggested relevant queries based on search volume (Google Support n.d.).

[38] This choice was based on a comparative analysis between #pakistan, #pkflood, #pkrelief, #pkfloods and #helpPakistan, using the WTHashtag service (Hootsuite Media n.d.) to find the dominant hashtag as adapted by users.

[39] Platforms rename features over time, for example Posts by Everyone is now called Public Posts, but throughout the dissertation I use the names of platform features as they were at the time of the case study.

by relevance, by date or by some variation on these.[40] In this case study we looked at the two dominant modes of organizing content, relevance and freshness, which are shared by most of the selected digital media. For the eight media the default mode for presenting results was: Facebook (posts) by date, Google News by date, Wikipedia by date, Google by relevance, Google Blog Search by relevance, YouTube by relevance, Flickr by relevance and Twitter also by relevance (displaying Top Tweets). These default modes show that the search output of devices does not necessarily reflect how 'Web 2.0's interfaces heavily emphasize the new even at the cost of other modes of organization such as relevance or importance' (Gehl 2011, 5); at the time of the case study Twitter emphasizes fresh content in its default interface, but they turn to relevance as their mode for presenting content as the output of a search query.

Before querying and saving the results a 'research browser' was created and the interface settings of each of the devices were adjusted to meet the research design, that is, the result pages were ordered by freshness and relevance if possible.[41] Following the recommendations made by the devices in the query design, we chose to use the default settings for result pages if possible (for example with the Google engines, this meant retaining the default 10 results per page). From each of these devices, the content of the page was saved on a local computer with an interval of five minutes and the number of new results compared to the previous interval was noted. Finally, the number of new updates or new results within the relevance setting for each interval was calculated for each platform or engine and plotted in a barcode chart, each barcode line representing one new piece of content (see Figure 8).

Figure 9 shows the results of the study. For each device at least one, if not two barcode, charts of its specific pace are shown. The charts on the left show the pace based on content sorted by freshness; the charts on the right are based on relevance. The patterns that unfold when looking at the barcode charts from left to right show the pace of new content across an interval of 24 hours. The legend on the bottom gives an indication of the time (CET) at which change happens in the interface. With each line symbolizing a new entity of content for the query issued, it becomes apparent that digital media have their own pace, which differs in terms of intensity (number of lines), rhythm (pattern of lines occurring) and variation over time.

[40] For example, Flickr's default result page offers results sorted by 'Relevant', 'Recent' and 'Interesting' the latter also taking user activities such as commenting or favoring into account (Flickr n.d.).

[41] See (Rogers 2013b) and Chapter 5 for a more elaborate explanation of the notion of a research browser.

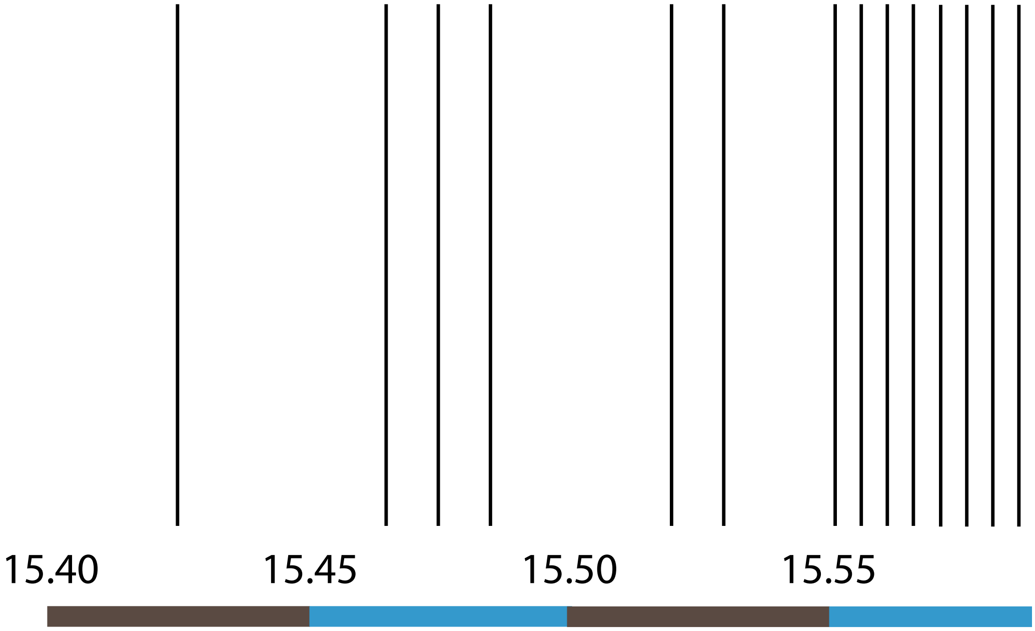


Figure 8: Barcode chart showing the pace of new content. Each line represents a new result. Lines are distributed over the five minute intervals, the denser the lines within an interval, the more new results compared to the previous five minute interval. Visualization created by using custom code and Adobe Illustrator, 2010. Source: *Borra et al. 2010*.

The visualization points to two broader findings. First, the web appears to have a series of distinctive 'realtime cultures'. As argued previously, digital media tend to follow specific update cycles. But the realtime cultures of platforms and engines are not only formatted by the flow of fresh content to be processed in the back-end, they are also shaped by how the devices offer content in the front-end through their interfaces. By comparing relevance with freshness, we can start appreciating how web devices construct their own specific realtimes. It is not surprising that the 'relevance' mode seems to slow down the pace of content compared to freshness, as it filters through social and cultural cues such as authority of source or uptake, and commercial ones such as sponsored results. The exception is Flickr, where in the default relevance mode new photos are shown each time the page is refreshed.

Second, the findings of the case study suggest that pace indeed has different patterns. Pace is not only connected to the digital medium's algorithmic logic but also to different types of web activities such as posting, linking, editing, sharing, uploading, and retweeting connected to each medium. Platforms and engines allow for different

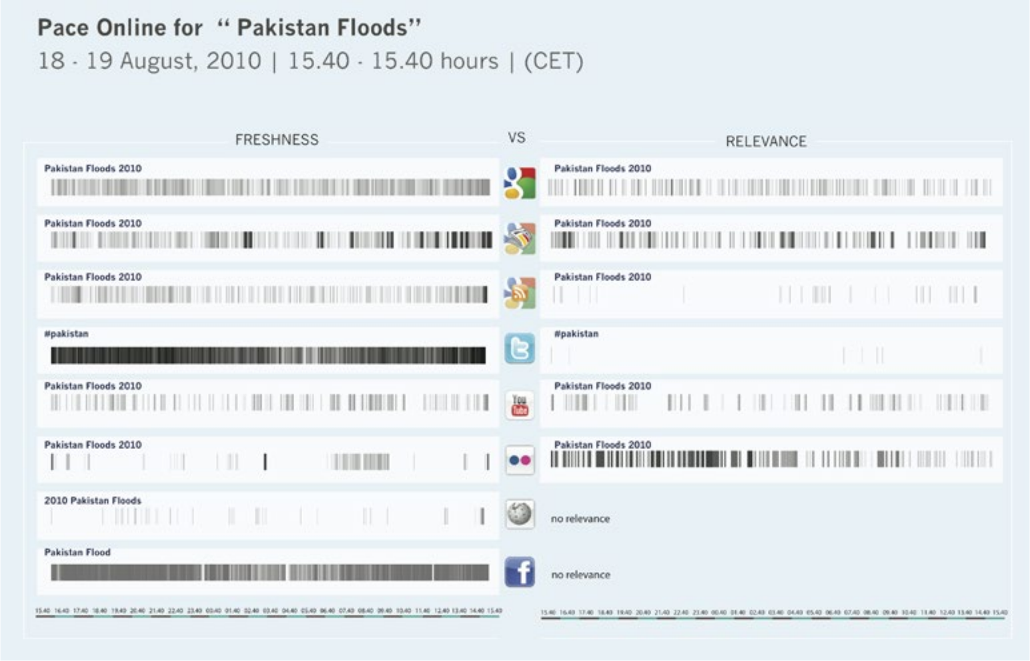


Figure 9: Pace Online for "Pakistan Floods". Barcode chart showing the pace of content for the issue Pakistan Floods in Google Web, Google News, Google Blogs, Twitter, YouTube, Flickr, Wikipedia and Facebook in both freshness and relevance mode, 18-19 August 2010. Each line represents a new result. Lines are distributed over the five minute intervals, the denser the lines, the more new results compared to the previous five minute interval. Facebook's pace is determined solely by freshness, while for example Google News' pace is based both on freshness and relevance. Visualization created by using custom code and Adobe Illustrator, 2010. Source: *Borra et al. 2010*.

activities that contribute to the specific pace of a medium. On the micro-blogging platform Twitter the pace includes platform activities such as tweeting and retweeting; on average 5,700 tweets are uploaded every second (Krikorian 2013), or 342,000 every minute. Compared with this overall rate of activity on Twitter, the rate for [#pakistan] shows an average of 5.84 new results per minute over the 24 hours collected in this case study. Between 2 PM CET and 3 PM CET, Twitter reaches its peak with an average of 16.4 new results per minute.

Analyzing the barcode charts allows for defining three distinct patterns of pace: stream pace, bulk pace and stale pace. Stream pace is continuous and frequent. The platforms that display a typical stream pace request users to add new content, such as Twitter and Facebook. Google, in freshness mode, also displays characteristics of stream pace as it shows the number of fresh results indexed by the engine. Looking more closely at the prototypical stream platforms Twitter and Facebook, we found

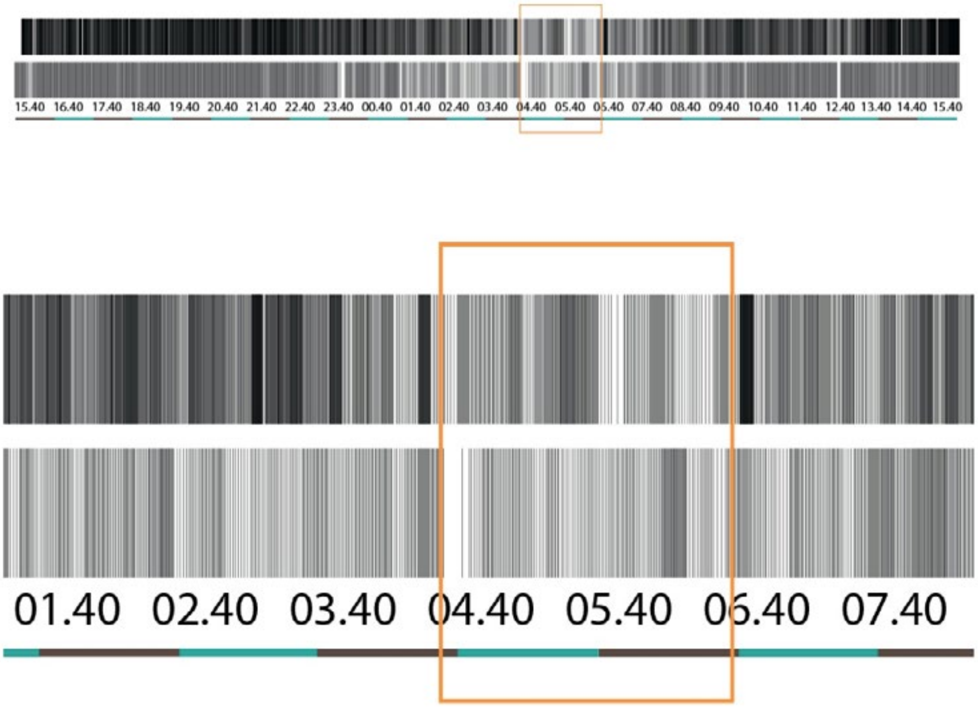


Figure 10: Barcode chart showing the pace of new content for the issue 'Pakistan Floods' on Twitter, 19 August 2010. Each line represents a new result. Lines are distributed over the five minute intervals, the denser the lines, the more new results compared to the previous five minute interval. The orange square indicates idle time on Twitter. Visualization created by using custom code and Adobe Illustrator, 2010. Source: *Borra et al. 2010.*

that the pace of fresh content is indeed high and continuous. The results, however, also show a small decline in activity around 5 AM CET where the stream is idle, indicating some form of day and night rhythm in Twitter use (see Figure 10). Second, bulk pace is characterized by fresh content being added in larger or smaller bulks. Typically this type of pace is found in media where users upload content in batches, such as user-generated content platforms such as Flickr and to a lesser extent YouTube and Google Blogs. The same is noticed in Google News where editors follow news cycles or in Wikipedia where edits tend to be done in batches as users often save multiple minor edits to a single article within a short timeframe. Some media offer new content in bulk pace in relevance mode, too, such as Flickr, Google News, YouTube and to a lesser extent Google. Third, the pattern of stale pace directly refers to the lack of frequent new content and specifically to those media that hardly change their results compared to the others. Google Blogs does not update its relevant stories very often.

In comparison, the rate of fresh content in Google in its default relevance mode is rather fast. Moreover, Twitter, which in relevance mode displays 'Top Tweets' at the top of the result page, is rather stable and shows a very slow pace.

From realtime to realtimeness

The empirical study opens up a series of contributions on the realtime web and subsequently on medium research. To start with, it permits an account of the *making of realtime* which does not unfold as a flat, eternal now or as a global, high-paced stream, but points to particular web specific entities, activities and actors determining the temporality of the specific space, as the issue studied unfolds at a different speed in relation to different media. I have tried to demonstrate how the patterns of pace are specific to digital media, and are the outcome of the interplay between content, its storing and algorithmic processing, interfaces, search and rank algorithms, queries, user activities, but also time and date. After tracing the specific elements that are involved in the assembly of a device-specific realtime, the case study also pointed to the multiplicity of realtime. The different paces detected in the barcode chart—stream, bulk or stale pace—hint at the multiplicity of rhythm. The pace of freshness indicates the rate at which new content in relation to the query is being produced and presented in platforms or taken into account in engine results, while the pace according to relevance is a mode of temporality determined by the devices' algorithmic relevance. Freshness and relevance emerge as two modes of organizing the process of pacing growing amounts of information, and therefore as internally different ways of introducing rhythm and pattern to their circulation.

When further exploring the different devices of the study and their interfaces, more features and practices contribute to the multiplication of realtime. In fact, the dominant mode of presentation implies that no single 'generic' or global stream is experienced by all users in a similar way (Manovich 2012c);[42] in Twitter users can no longer follow the general 'all tweets' stream and in Facebook a Public News Feed containing all public messages has never existed.[43] Users have to assemble their own streams based on followings, friend connections, hashtags or News Feed settings,

[42] The implications of which for using the output of devices for digital social research are further discussed in Chapter 5.

[43] Also developers cannot get access to the full stream of tweets, called the firehose, but instead access is licensed to re-sellers such as Topsy, Gnip and DataSift or through commercial partnerships. Alternatively, developers and researchers can use the 1% or 10% samples provided by Twitter.

which all come with their specific pace.[44] Many realtime devices, most notably those with streams, also bring along a series of third-party mediators or clients that enable users to view content at a different pace by reassembling platform content into new forms. In the case of Twitter, the platform itself recently introduced 'custom timelines', allowing users to assemble their own streams (Ellin 2013); Tweetdeck allows users to configure multiple custom timelines in columns arranged side-by-side, each coming with their own distinct pace. Hence, what users are dealing with are ecologies of streams (Berry 2011c), as the same data may be assembled into different streams.

Realtime is thus not only fabricated differently by devices, but also specific and multiple in their respective front- and back-ends. As introduced before, scholars such as Gehl (2011) have outlined the coupling between the realtime drive in the front-end and archival impulse in the back-end. What emerges in the context of this empirical work on pace, however, is an insight in the interconnection of the front- and back-end temporality, as well as a further differentiation of this dual temporality. From a device-driven approach, the experience of realtime in the front-end of digital media is not always supported by visual cues. For example, while the pace barcode charts feature the change rate of search results presented in the front-end of search engines, these search engines typically do not present visual cues for the pace of new content unfolding to web users. In contrast, the web version of Twitter does indicate 'x new Tweets' on the top of the stream. Both confirm and extend the universal temporal regime (Lovink 2012) proposed by Gehl (2011), as what emerges is a perspective on the medium-specificity of the realtime experience.

Whilst many critics of realtime media and research have outlined the encapsulation into the present (Uprichard 2012), a more complex simultaneity and folding of temporalities is also at stake. The fabrication of realtime may entail the interplay between past, present and future as studied in the project Historical Controversies Now (Figure 4) in Chapter 2. For instance, in order to feature high on the relevance based search results, Google's PageRank Algorithm determines the relevance of results based on the relative authority of the source, taking into account past links, recommendations and content organization, as will be elaborated upon in Chapter 5. In the case of Twitter, which simultaneously displays fresh, new content and relevant, featured results, relevance becomes a recommendation feature that alters the pace of the freshness stream, as these so-called Top Tweets are designed to produce future user engagement by making them sticky and stay on top of a fast changing stream. Relevance thus brings the past of sources or web users together with a potential future, creating a multi-layered account of realtime. This quality was already noted by

[44] There are exceptions; for example, the ad-free social platform App.net provides a global stream.

Berry, who claims: ‘To be computable, the stream must be inscribed, written down, or recorded, and then it can be endlessly recombined, disseminated, processed and computed’ (2011b, 151). A stream is therefore not just the inflow of new content, but also its constant recombination or pacing based on algorithms, featured content and user activities.

Here I return to the question of political economies built into the fabrication of realtime. Features like recommended Tweets allow companies or users to alter the fast paced temporality of Twitter streams by giving content duration, pacing it differently and making it sticky by paying money.[45] Similarly, Facebook offers various related features that allow pacing down the stream against payment, including promoted posts, recommendations and featured pages. Underneath their posts, users are offered the possibility to ‘promote’ their content so it will remain on the top of their friends’ Newsfeed for a longer period of time and will thus receive more attention and interaction. Such promoted content contributes to direct the immediacy of user engagement towards specific content, introducing the steering of engagement with the stream, for instance through the grammar of liking (Gerlitz and Helmond 2013).

In a similar fashion, social media marketing deals with creating content in such a way that users keep sharing, so that it returns and circulates across as many Newsfeeds as possible. Again, the objective is to slow down the disappearance of content in the fast paced stream whilst increasing the pace of its interaction; the reduction and speeding up of the pace of different actions are thus tied together. To reinforce this process, a number of third-party services built on top of social media platforms offer scheduling services that can post content at set times to pace the distribution of their content and ensure it will enter the stream at a strategically relevant time. For Twitter such services include Hootsuite, which contains auto-scheduling features, and Buffer, which uses the Tweriod algorithm to find the optimal moment to post and schedule tweets (Widrich 2012). Whilst platforms may have a general interest in high paced content production and user engagement, as argued by Gehl (2011), together with the numerous third party applications they also cater for different temporal interests of cooperating partners and paying clients who want to slow down the pace of streams or introduce selected sticky content to gain attention.

Considering this specific fabrication of times in devices, I suggest to think about time not as an event happening *in* realtime, in the now, but as entangled in the fabrication of specific forms of realtime*ness*. Realtimeness is the continuous movement of new content, its request and display in devices, as well as the engagement by users through

[45] Also see the work of Taina Bucher, who is developing the notion of the ‘right time’ (Bucher 2014).

web activities and the filtering of content based on freshness and relevance. In this sense, realtimeness refers to an understanding of time that is embedded in and immanent to platforms, engines and their cultures. Pursuing the idea of such immanent and device-specific time further, realtimeness underlines how the specificity of time cannot be accounted for from the outside, applying extraneous measures, but only from the inside, tracing the increasing or decreasing intensity of pace in each digital medium and its internal variation. Realtimeness unfolds as a temporal condition to which web users have to respond; it can—as shown in the pace research—itself be subject to variation and it is assembled from the inside of medium engagement. Most notably, it is thus more specific and immanent than Leong et al's (2009) notion of the multiplicity of time, as it also makes the front- and back-end temporalities of devices act together. Not only does it specify the more general account of the eternal now and the realtime web, but it also allows for an empirical perspective on what Leong et al proclaim to be multiple times.

Conclusion

By developing this multiple and empirical account of realtime(ness), the chapter seeks to contribute to a variety of discourses in relation to the growing interest in realtime media. Firstly, the notion of pace permits to consider the fabrication of realtime as managing the constant and dynamic production of content. In the context of devices, the continuity of content production is a fundamental building block of engines and platforms, which rely on the constant provision of new content and interactions. The organization of the pace of updates can be considered as a pattern through which the continuous production of new content is being organized in ways that are aligned with the specific politics of digital media. Beyond my focus on pace as descriptor of fabricated duration or realtime, further descriptors might contribute to the qualification and specification of realtime, such as duration, stickiness of content, volume of data, halftime or speed. Whilst future research has to explain these descriptors, the study of pace allows for a more multi-layered notion of computational realtime and offers alternatives to the generalized use of realtime as a marketing buzz term.

Secondly, the empirical account on realtimeness differs from a variety of current interests in realtime research. In view of a growing interest in tracing and analyzing phenomena as they happen, social researchers have recently invested in developing methods for live or realtime research (Back and Puwar 2012; Back, Lury, and Zimmer 2013; Elmer 2013). Although the approach in this chapter was also concerned with the happening of content, it specifically focused on the socio-technical conditions in which digital media create such live content and how this may intervene in the method. Instead of focusing on making research itself 'live', the making of what constitutes the 'live itself' should be focused on. This approach thus shared an interest in the conditions of making of data in live time, such as developed in Chapter 2. There I also emphasized temporality as internal to digital media and suggested a form of live research deploying the formats and life cycles of digital data as formatted through devices for analysis. In this chapter, I am more concerned with differentiating live or realtime in itself and thus underline that such liveness cannot just be proclaimed; it has to be accomplished in the first place. Especially when taking the multiple temporalities which may entail a complex folding of past, present and futures in realtime media such as Facebook streams into account, it becomes apparent that conceptualizing realtime media as a mere expansion of the present misses the intricate folding of past, present and future and the operations built into these specific fabrications of realtime.

Third, the chapter complicates the relation between computational and experienced realtime. Although reflecting on the role of users, starting from the interplay between push and pull to the assembling of realtimeness, the chapter diverts from a purely phenomenological perspective to the experience of realtime. In previous debates realtime media and especially streams have often been described in terms of messianic, unrepresentable accounts of information (Berry 2011c). Such a perspective may coincide with the starting point that realtime media are organizers (or pacers) of information, which currently appears in 'gigantic' or in 'sheer unrepresentable' or even 'messianic' quantity. Engaging with realtime media means engaging with the expectation of new content, new users and new activities, which, anyhow remain ungraspable, unknown and push the limits of representability. In this chapter, rather than just focusing on such opaque and totalizing realtime experiences operated by global streams or black-boxed algorithms, my engagement with realtime technicity and the study of pace has shown that the potential infinity of new content can be further specified when considering realtime as immanent to and co-constituted by media device and their cultures. Especially social media platforms invest in a multiplicity of realtime features, which explicitly offer different paces, rhythms, and durations of content engagement to cater for the interests of their multiple cooperating partners. The realtimeness of different digital devices might be continuous and not limited by the end of new content production; yet it is fabricated and enacted in specific ways. Rather than perceiving platforms and engines as messianic media, this chapter suggests to consider them as pacers of realtimeness, whose specific pace is closely tied to the politics of such devices.

As shown in this chapter, medium research configures digital media devices in the research apparatus and operationalizes the media to make findings about those media. By studying the media and their device cultures themselves, this chapter contributes to software studies in the sense that it not only inspects the settings and interfaces of the media but repurposes the operational capacities of digital media and uses the data that pass through it as material to study the media in action, as well as the way in which the data are organized and recommended. Such an approach thus allows one to empirically study what and how data is captured, as well as what it affords for digital research, which will be the topic of the next chapters. Apart from the theoretical and empirical perspective on realtime, this chapter has moreover provided a historical periodization of dominant knowledge logics dealing with realtime. Digital cultures, in this respect, have moved from directories and portals, to new dominant knowledge logics such as (fresh) search engines and streams. As such, this chapter pleas for a constant re-evaluation of what it is that we are studying, both in terms of its medium specificity, as well as how that affords different modes of digital research.

Conjuring up a past state of the Dutch blogosphere

Chapter 4

IN THIS and the following chapter I engage with the volatility of method by focusing on how specific configurations of digital devices and digital method may affect the behavior of their components. The previous two chapters discussed the perspectives that are key to digital research: the social research perspective in Chapter 2 and the medium research perspective in Chapter 3. When focusing on the volatility of methods, I examine the flexibility and indeterminacy of the various components of the methodological apparatus in this chapter and the update culture of digital media in the next chapter. In digital research the volatility of digital methods signals the intricate relation between medium and method; in the method the research aim is negotiated through the device research affordances and cultures of use. Digital research is therefore always in part medium research, as it highlights the specific purposes and use practices inscribed into digital devices and the consequences of specific configurations on the findings. In this chapter I engage with volatility by inquiring into the evolution of features and platform software over time, and into the research modes afforded by the medium in the context of the research aim. The research affordances of the Wayback Machine are typically considered to be URL histories, web experience (not 404) and evidentiary functions (Rogers 2013b). The question then is what other modes of research besides the single site history the interface of the Wayback Machine affords.

The blogosphere played an instrumental role in the transition and evolution of linkig technologies and practices, such as the introduction and development of trackback, pingback and RSS. Blogging thus became a distinct digital culture. Important research in this area has been practice, event or issue based, trying to capture an otherwise fleeting phenomenon in realtime before it is deleted, overwritten or no longer available. In the previous chapters ramifications for such realtime research were addressed in detail; this chapter investigates the extent to which it is possible to do historical research on an object that is no longer, or only very partially, 'live' and available on the web.[46] Now that the blogosphere has reached maturity, the first historical accounts are created. This study seeks to contribute to this body of literature by investigating the blogosphere's platform and software infrastructure. More specifically, the empirical research focuses on mapping transitions over a certain period of time in a national blogosphere. The evolution of technologies in the Dutch blogosphere may also be approached in terms of the transitions in the 'grammars of action' encoded into the software that drives the blogosphere (Agre 1994). Transitions in linking technologies and practices leading to the reconstruction of and transitions in the Dutch blogosphere from 1999 to 2009 are mapped and traced. Longitudinal network analysis with archived data allows for the study of the rise and fall of blog platforms and

[46] See also Chapter 6 on Wikipedia, a platform logging its own history. This chapter researches the Dutch blogosphere, which has no archiving mechanisms of its own.

social media platforms within the blogosphere as well as for the investigation of local blog cultures. Research on historical data, however, also has to deal with the volatility of the medium, such as incomplete data sets and data loss.

Taking a device-driven perspective, the contribution will be both methodological and conceptual, as choices in configuring the method in relation to the digital medium shape the definition of blogs and blogosphere. This chapter addresses methodological questions related to the empirical research of a national historical blogosphere, and presents the outcome of the research into the Dutch blogosphere.[47] The proposed approach combines techniques that are also used by search engines and web archive crawlers with editorial techniques commonly used in social research. Although preliminary, the research both acts as a proof of concept and as a model for studying national and historical link networks such as blogospheres, and provides new insights into the shape of the Dutch blogosphere and its interconnections.

The blogosphere is often studied by mapping and visualizing the interconnections between blogs, in order to make it tangible and visible. In other words, the image of the blogosphere must be construed, either by blogosphere related services, such as directories, web rings and blog search or by academic network visualizations. By means of similar techniques, such as contemporary blog related services, network visualizations may be constructed; RSS feed-crawlers fetch the content—current and newly updated blog posts and their links—of blogs using their feeds (Bross et al. 2010) or use web crawlers for network analysis. Crawling and network analysis may be used for a wide variety of analytical purposes, such as the Issue Crawler for issue network analysis to track conversation patterns in the blogosphere (Bruns 2007); crawling the front page of blogs to reflect blogroll communities (Adamic and Glance 2005); and large scale grouping of linked blogs to define clusters of shared informational worlds (J. Kelly and Etling 2008). Although different tools and methods produce different network visualizations, they all provide a graphical representation of interconnections and insights into the overall structure of the blogosphere and its actors (Highfield 2009). This chapter examines how choices in method do not only shape the blogosphere but also shape the definition of the blogosphere and blogs.

Historical blogosphere research mainly consists of ethnographic research, providing personal stories and anecdotes (Blood 2004; Rosenberg 2009), as well as empirical work, such as Michael Stevenson's (2010) research on the early A-list blogosphere, Rudolf Ammann's (2009) research project on the birth of the blogosphere and Ravi Kumar et al. (2004) researching the structure and evolution of LiveJournal blog space. Kumar et al. suggest a method based on time stamps—in addition to other features

[47] The empirical research was done in collaboration with Anne Helmond and Erik Borra.

such as those present in profile pages—to map a blog space over time. More generally, this research contributes to the growing interest in the history of the web and historical network analysis (Brügger 2009; Brügger 2010; Stevenson 2013; Ben-David and Huurdeman 2014; Liu 2011). Most notably the Internet Archive enables the study of previous states of the web by providing time–stamped snapshots. Although single-site history is preferred—only single URLs can be retrieved—Internet Archive data may be used in a variety of ways. Ammann (2009) studies the emerging blogosphere by mapping linking patterns of early blogs on the basis of the Internet Archive; Stevenson outlines a method to repurpose the Internet Archive to create a custom archive by using the early blog index EatonWeb as a historical resource to recreate the blogosphere. The research reported on in this chapter builds on the above-mentioned methods and tools and develops a number of novel techniques and methods.

First the historical blogosphere is studied by making snapshots of the Dutch blogosphere, specifically paying attention to actor definitions and interlinking practices by introducing fine-grained URL and source code analysis. Next it is endeavored to widen the historical blogosphere analysis beyond the Anglo-American context by specifically focusing on the Dutch blogosphere. By investigating the Dutch character of top-level domain, blog software and platform use the definition of a 'national blogosphere' is proposed. Finally, I aim to contribute to hyperlink network and issue network analysis research by redefining what is considered an actor in the blogosphere.

Dutch blogs in transition

How can the nationality of an otherwise 'placeless' digital site be formally defined? Anticipating the work in Chapter 7 on mapping the Iranian web, the web archiving community often tried to answer this question with locative technical indicators such as the IP address or top level domain (TLD). These, however, are always ambiguous and their usefulness highly depends on their purpose and application. In view of saving digital heritage for posterity, the Dutch web archiving institution formulated three defining characteristics, namely language, TLD, and subject matter 'aimed at the Netherlands' (in Weltevrede 2009) which are rather difficult to automate and will be discussed in more detail in Chapter 7. First authoritative sources with their selection criteria for including blogs in their lists were chosen to define a Dutch blog in this research project. In a second step, the evolution of the blogging software and features were analyzed.

The collection of blogs in the corpus was retrieved from a 2001 database dump—containing 631 unique blogs—from Loglijst, an early Dutch blogosphere indexing initiative. In addition, expert lists were compiled from interviews, books and authoritative lists found on the web and in the Internet Archive. These expert lists included long list nominations for the Dutch blog awards, the Dutch Bloggies from 2001-2008, all blogs mentioned in two seminal pieces on the history of the Dutch blogosphere by historians Frank Schaap (2004) and Frank Meeuwsen (2010) and finally a list citing 'Weblogs that really matter' in a December 2010 blog post by Bert Brussen, blogger for the famous Dutch 'shocklog' Geenstijl (2010). Relying on these sources to assemble a collection of Dutch blogs led to the inclusion of a small number of Belgian (Dutch language) blogs that these sources considered to be part of the Dutch blogosphere.[48]

Subsequently, the Internet Archive's new Wayback Machine was queried for each blog's URL and the result selected was dated closest to the middle of each investigated year. From the 2,507 URLs requested 946 blogs could be retrieved from the Internet Archive. This method yielded a collection of archived copies of historical Dutch blogs for each year with a timestamp near the middle of the year. Only blogs with a copy in the Internet Archive were retained for further analysis. Table 1 depicts the number of blogs per year serving as a starting point.

1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
24	138	456	816	778	863	850	788	717	860	723

Table 1: The number of blogs from the expert list that were available in the Internet Archive Wayback Machine, per year from 1999 until 2009. The URLs were retrieved with the Internet Archive Wayback Machine Network Per Year tool, 2011. Source: Weltevrede and Helmond 2012a.

Blog software features have previously been studied in relation to the popularity and success of weblogs (Du and Wagner 2006) and in relation to blogging practices enabling or restricting certain actions (J. Schmidt 2007). In addition, this study interrogated the starting points for evolutions in the TLDs, platforms and software used. This study focuses on digital culture—national digital culture—by investigating a further specificity in Dutch digital practice, that is, software and platform usage as well as persisting applications, despite Twitter, Technorati and other dominant services from the U.S. Therefore specific attention was paid to the origin of the software and platforms Dutch bloggers talk about when describing Dutch blogging practices. The inquiry is three-fold and includes the analysis of TLDs, platforms and self–hosted blog software.

[48] Belgium borders the Netherlands and the countries share a common language: Dutch.

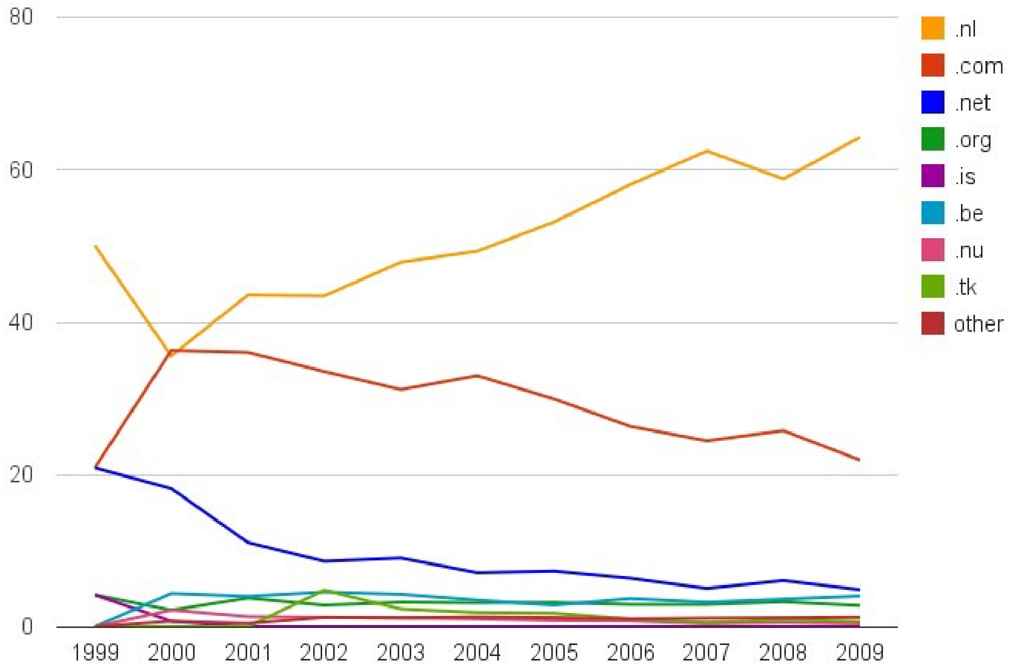


Figure 11: Relative distribution of Top Level Domains (TLDs) in the Dutch blogs from 1999 until 2009. The Dutch blogs in the collection retrieved from the Wayback Machine favor the .nl domain over all other domains throughout the years. Moreover, a significant increase in the .nl domain becomes apparent, whereas the .com domain is steadily losing share over time. Visualization created in Google Spreadsheets, 2011. Source: Weltevrede and Helmond 2012a.

TLD analysis

The top-level domain (TLD) analysis presented here is part of a larger series of URL analysis methods discussed in this chapter. As a first step the TLDs of the starting points were counted by entering URLs in batches corresponding to a single year by means the TLD Count tool. Figure 11 shows the relative distribution of TLD usage over time. The Dutch blogs in the collection favor the .nl domain over all other domains throughout the years. Moreover, the increase in the .nl domain is significant, whereas the .com domain is steadily losing share over time; the results in the next section show that Dutch bloggers move away from .com blogging platforms such as Blogger's Blogspot and go to Dutch .nl blogging platforms.

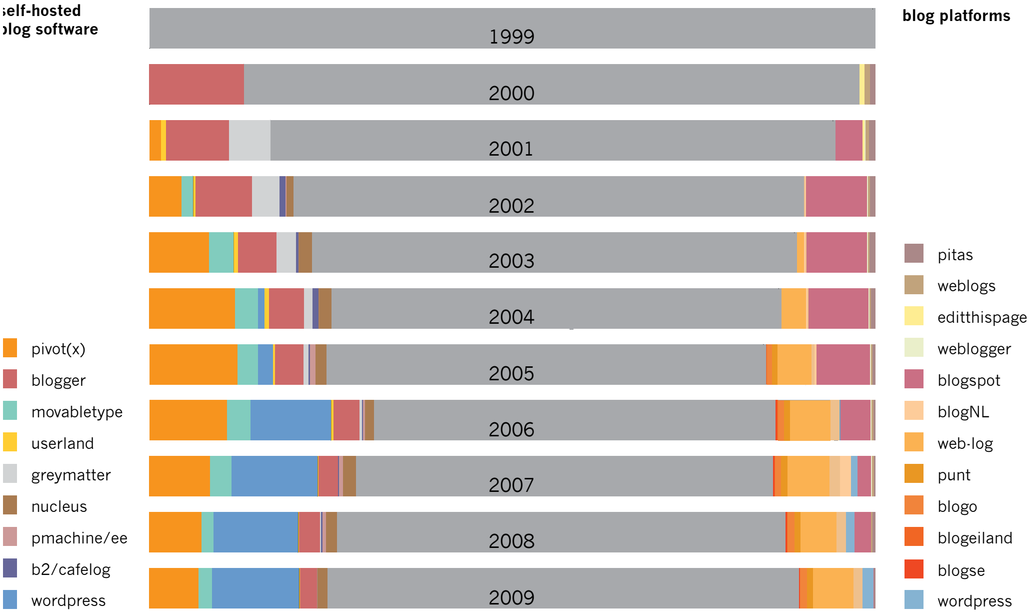


Figure 12: Relative distribution of self–hosted blog software & blog platforms in Dutch blogs from 1999 until 2009. The graph shows the rise and popularity of Blogger's platform, Blogspot, in the beginning of 2000 in the collection of blogs retrieved from the Wayback Machine. The decline of Blogspot coincides with the rise of the Web–Log.nl blogging platform, and other Dutch blog platforms such as BlogNL, Blogo, Blogse, Punt and Blogeiland. The figure clearly shows how from 2004–2005 onwards Dutch bloggers—except for a relatively small number of Blogspot and WordPress.com users—shift to Dutch platforms, which are orange color–coded. Only a few bloggers remain on legacy platforms such as Pitas, which no longer accept new members but are still functional for old members. Visualization created in Google Spreadsheets and Illustrator, 2011. Source: Weltevrede and Helmond 2012a.

The Dutch .nl domain is one of the top five country code top-level domains (ccTLDs) in the world, which is also reflected in the Dutch blogs (SIDN 2010). It is however remarkable that the .nl domain has been dominant from the beginning, although the .nl domain only became available to private individuals in 2003. As a forerunner, since 2000 individuals were allowed to register third-level domains such as jansen.123.nl but these domains were rather rare and do not appear in our collection of blogs (SIDN 2007). As stated before, the Dutch blog collection contains a number of .be blogs, steady from 2000 onwards. Furthermore, 2002 presents a peak of .tk. Dot.tk 'Renaming the Internet' offers free domain names and includes URL redirection and forwarding services. Lastly, a number of domains is unconventionally used for 'commercial or vanity' purposes, including .nu (country code for Niue), marketed as 'now' in Dutch and .is (country code for Iceland), which in Dutch is the third person singular of the verb 'to be' (Wikipedia contributors 2012a).

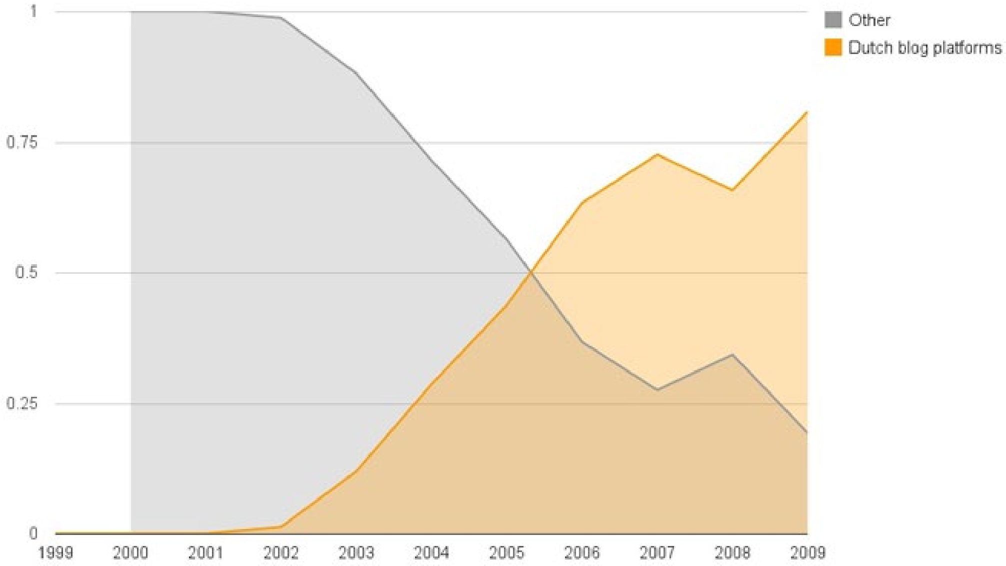


Figure 13: The relative amount of Dutch blog platforms over time compared to other blog platforms from 1999 until 2009. In 2009 almost all bloggers on blog platforms make use of Dutch platforms in the collection of blogs retrieved from the Wayback Machine. Visualization created in Google Spreadsheets, 2011. Source: Weltevrede and Helmond 2012a.

Platform analysis

A second analysis complementing the TLD analysis visualizes the variety and proportion of blog platforms used in the Dutch blogosphere. This requires basic background knowledge of blog platforms. With the use of Google Refine, 'a power tool for working with messy data'[49] each of the blog platforms was 'coded' in GREL (Google Refine Expression Language) to automatically search, transform and count the platforms in the set of URLs. The results are presented in Figure 12, a custom made visualization combining the blog platform analysis with the self-hosted software analysis as discussed in the next section.

The graph shows the rise and popularity of Blogger's platform, Blogspot, in the beginning of 2000. The decline of Blogspot coincides with the rise of the Web-Log.nl

[49] Originally created for Metaweb Technologies, Inc. by David Huynh. Metaweb Technologies, Inc. was acquired by Google, Inc. in July 2010 after which it was named Google Refine. After Google discontinued the development it was released as open source software in October 2012 under the name Open Refine.

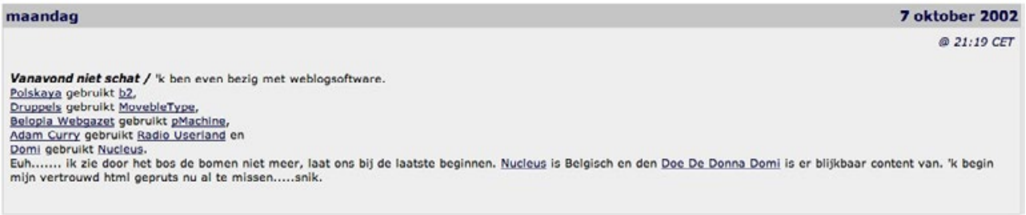
blogging platform, and other Dutch blog platforms such as BlogNL, Blogo, Blogse, Punt and Blogeiland. Figure 12 clearly shows how from 2004 and 2005 onwards, Dutch bloggers—except for a relatively small number of Blogspot and WordPress.com users—shifted to Dutch platforms, which are orange color-coded. Only a few bloggers remained on legacy platforms such as Pitas, which no longer accepted new members but were still functional for old members. Dutch software and platforms played an important role in the Dutch blogosphere and between 2004 and 2009 over 40 percent of all bloggers used Dutch blog software or Dutch blog platforms. When zooming into the use of platforms only, in 2009 almost all bloggers on blog platforms made use of Dutch platforms (see Figure 13).[50]

Self–hosted software analysis

URLs were analyzed to investigate the distribution of TLDs and platforms used in the Dutch blogosphere. The outcome suggests that the early Dutch bloggers did not use blog platforms. In general, they preferred to manually create their blogs, written in HTML or they used specifically designed self–hosted blog software. In HTML, the reverse-chronology, which is considered to be a key characteristic of blogs (Blood 2004; boyd 2006), had to be manually enforced in order to place the latest blog post on top. In order to include these kinds of blogs in the analysis a method going beyond the blog's URL had to be developed, to search within the page's source code for the URL referencing the software powering the blog to create an accurate list of blog software.

Initially, the list was compiled by analyzing maps, which I return to shortly, and then refined with newly discovered blog software throughout the research project. To compile the list of self-hosting software, the reflexivity of bloggers was used. Typically, bloggers tended to analyze and describe the practice of blogging (Hourihan 2002; Blood 2004). When researching the initial list of software, a number of blog posts were discovered comparing or mentioning different types of software (for an example, see Figure 14). For each year the source code of the collection of archived blog front pages was searched for the presence of the blog software types with the Source Code Search tool. The results were editorially checked to establish whether the reference to the software implied that the blog was indeed running on it. Especially in the beginning, references to self-hosted blog software were not standardized. In later years the 'powered by' button in the side bar or footer became standard for most self-hosting software.

[50] In 1999, none of the blogs were located on blog platforms because the first platforms were only introduced around this time. For example, Pitas (later known as Blogger) was created in July 1999, Diaryland in September 1999 (Helmond 2008).



maandag **7 oktober 2002**

@ 21:19 CET

Vanavond niet schat / 'k ben even bezig met weblogsoftware.
Polskaya gebruikt b2,
Druppels gebruikt MovebleType,
Belopia Webgazet gebruikt pMachine,
Adam Curry gebruikt Radio Userland en
Domi gebruikt Nucleus.
Euh....... ik zie door het bos de bomen niet meer, laat ons bij de laatste beginnen. Nucleus is Belgisch en den Doe De Donna Domi is er blijkbaar content van. 'k begin mijn vertrouwd html gepruts nu al te missen.....snik.

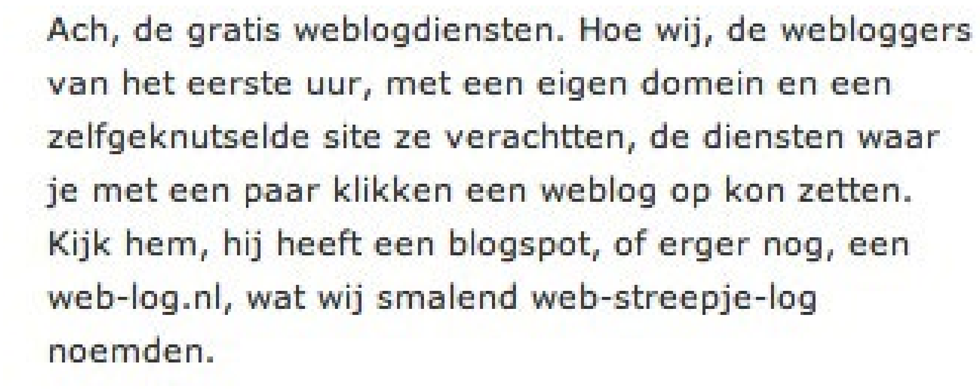
Figure 14: ‘Vanavond Niet Schat’. Partial screenshot of blog post from the Dutch Blogosphere on 7 October 2002. The weblog author writes ‘Vanavond niet schat / ‘k ben even bezig met weblogsoftware’ which translates to ‘Not tonight love/I’m busy playing around with weblog software’. Blog post retrieved via the Internet Archive Wayback Machine. Source: Vrints 2002.

Contrary to the blog platform counts, the self-hosted blog software results suggested that the Dutch blog software Pivot/PivotX had been powering Dutch blogs from the start, appearing to be the most frequently used software in the heydays of Dutch blogging. The decline of Blogger, the first blog platform used by Dutch bloggers, coincided with the rise of Blogspot—Blogger’s platform. Furthermore, the bar graph shows a boost of blogs powered by WordPress.org in the blogosphere from 2006 onwards. Movable Type and the Belgian Nucleus had a small but loyal share of bloggers running the software.

In terms of blog software and blog platforms, Dutch blogs peaked around 2005 for platforms and 2006 for software. Notably, the share of self-hosted software exceeded one-click publishing platforms, which even the bloggers themselves had not expected. A number of posts from early bloggers expressed fear that soon everybody would be blogging; some others voiced rivalry between self-hosting bloggers and platform bloggers (for an example, see Figure 15). Next this methodology was further developed by formalizing the various types of references to software, throughout the years, and design queries to automate the collection and analysis process of the results. The first part the study focused on designing methods to identify and map the evolution of the use of TLDs, platforms and self-hosted software driving the Dutch blogosphere. In the following part the interconnections between these blogs are examined as well as a method to create historical blogospheres.

Reconstructing the blogosphere

In 1999, Brad L. Graham coined the term ‘blogosphere’ to mark the end of cyberspace: ‘Goodbye, cyberspace! Hello, blogiverse! Blogosphere? Blogmos?’ (1999, n.d.). William Quick revived the word as ‘the intellectual cyberspace we bloggers occupy’,



Ach, de gratis weblogdiensten. Hoe wij, de webloggers van het eerste uur, met een eigen domein en een zelfgeknutselde site ze verachtten, de diensten waar je met een paar klikken een weblog op kon zetten. Kijk hem, hij heeft een blogspot, of erger nog, een web-log.nl, wat wij smalend web-streepje-log noemden.

Figure 15: Early Dutch blogger about the rise of free blog platforms. Source: Van den Berg 2009.

explicitly mentioning that the blogosphere is a space for serious discourse (2002). Echoing the idea of the blogosphere as a discursive space 'the imagined public sphere' (boyd 2006, n.d.)was presented alongside the idea of blogs as a reaction to mainstream media (Lovink 2007). Besides the notion of the blogosphere as a space for discourse, other definitions stress the formalistic characteristics of the blogosphere as an interlinked set of blogs which 'allows for the networked, decentralized, distributed discussion and deliberation on a wide range of topics' (Bruns, Kirchhoff, and Nicolai 2011, n.d.). Another approach to the blogosphere as an interlinked set of blogs looks at how blogs are 'embedded into a much bigger picture: a segmented and independent public that dynamically evolves and functions according to its own rules and with ever–changing protagonists, a network also known as the "blogosphere"' (Bross et al. 2010, n.d.). Following this line of thinking, the blogosphere may also be defined by including the actors they link to in their networked ecology: 'The notion of a mini-blogsphere additionally rests on the extent to which the set of blogs doing an issue are interconnected by links and/or by textual referencing', Richard Rogers continues, blogs also may 'be "connected" together through common references to a third party, for example, all blogs linking to or referencing a particular piece in *The New York Times'* (Rogers 2005, 7). Although these two dominant approaches to research the blogosphere have different objects of research, they do not exclude each other, as demonstrated by Benkler and Shaw's (2012) U.S. political blogosphere research.

Although highly formal, the blogosphere has more of a cultural than a technical meaning (Shaw and Benkler 2012, 472), because as illustrated in the previous section, the many different blog platforms and software types permit to custom use the blog software. At first glance the approach put forward in this chapter might appear formalistic, because the definition of the blogosphere follows from the outlined method based on link analysis (see below). By mapping the formal changes in linking patterns and URLs over time, however, findings about specific local cultures of use can be proposed.

The annual blogospheres are created from a collection of blogs retrieved from the Internet Archive by means of custom tools. One of the consequences of studying transition with the Internet Archive is that it enables only research on front-page level and not on a post level. Hence this method may be viewed as a structural 'blogosphere' analysis instead of an 'issue' or 'event' analysis. Although fully aware that the choice of starting points shapes the Dutch blogosphere, the methodology used only retains blogs deemed relevant by other blogs. It is a co-link analysis as used by the Issue Crawler, performed in two steps: first, for each blog all links on front-page level are extracted (one depth) and subsequently, in Gephi, only nodes receiving at least two links from the starting points are maintained in the network visualization (one iteration).[51] The resulting network map thus retains only co-linked actors, receiving at least two links from the starting points. This implies that the starting points themselves might drop off the map and that Dutch blogs, which are not available in the Internet Archive, might reappear in the network. It thus also acts as a validation of our expert lists.

Whereas the co-link analysis is most successful for locating issue networks, in this case, the specific configuration between mode of as and device results in issue or event-based links being excluded from analysis. In other words, in this case the alignment between the digital medium's research affordances and the mode of analysis leads to a different method, which has repercussions for interpreting the findings. This has three main reasons. First, the starting points are not chosen because they share an issue or an interest in an event, but rather because they share the practice of blogging in the Dutch web space. Second, only front-pages are crawled, which means that more structural links are followed, such as links in blog rolls, and to blog related services and blog software. In other words, these links are the stable variable in the analysis, whereas links in posts are only taken into account if present on the front-page. Third, the time frame of each network is one year. Combined with the previous point, that links from posts are only crawled one level deep, the result is that links to versatile issues dominating the Dutch blogosphere for a short period of time being are excluded, and only the more structural links prevail. When studying a structural blogosphere blogs are assumed to be embedded in a larger networked ecology created by bloggers through their linking practices, including other actors than blogs, such as blog portals, web rings, news websites and social media platforms.

[51] Issue Crawler is a software tool locating and visualizing networks on the web. Gephi is a software tool to explore and visualize networks.

In what follows, I will further describe how the Dutch blogosphere was re-constructed on the basis of pages from the Internet Archive and how the re-constructed blogosphere was prepared for further analysis. Specific attention will be paid to the construction process by reconfiguring actor definitions and reconsidering interlinking practices. This approach gives novel insights into the composition of the blogosphere and its actors. The reported methods were developed to study transitions in the historical blogosphere with the Internet Archive's Wayback Machine. The method consisted of two strands: first the network analysis was refined by defining the actors using Gephi and G–Atlas software, and subsequently the network analysis was complemented by color-coding the platforms present in the blogosphere.[52]

Defining the actors

As previously described, snapshots of the list of blogs from 1999 to 2009 were retrieved from the Internet Archive, extracting their outlinks on a front-page level and translating the results in Gephi's GEXF format. In Gephi, a simplified version of Issue Crawler's co-link analysis was performed so that only blogs with more than two links from our starting list were retained. Co-link was performed on a 'by site' level, as it was more indulgent than the 'by page' option because it counted all links from site to site. In other words, co-link analysis was performed on the hosts and not on the deep pages.

A common problem in digital network visualizations is that big platform nodes take a prominent position in the graph. Analysis of these maps often suggests that the debate is moving elsewhere (predominantly to social media). In an attempt to untangle the big platform nodes in the Dutch blogosphere the nodes of the network were redefined to actors . Most network analysis software treats the host and in some cases sub-host as the actor. However, the uses of URLs has evolved over time and in this case the 'actor' or blogger is often defined after the slash, like the early bloggers that started blogging from their personal homepage or the recent microbloggers on Twitter. A similar approach was developed by Médialab Sciences Po who defined the concept of 'web entities' to unravel pages grouped by domain name (Girard 2011), and was implemented in their network analysis software Hyphe (Médialab SciencesPo 2014). Also Benkler and Shaw (2012), in their work on the U.S. political blogosphere, stressed the importance to analyze what is inside the large network nodes in order to specify their internal differences.

[52] The G–Atlas software is developed by Mathieu Jacomy for TIC Migrations, Paris.

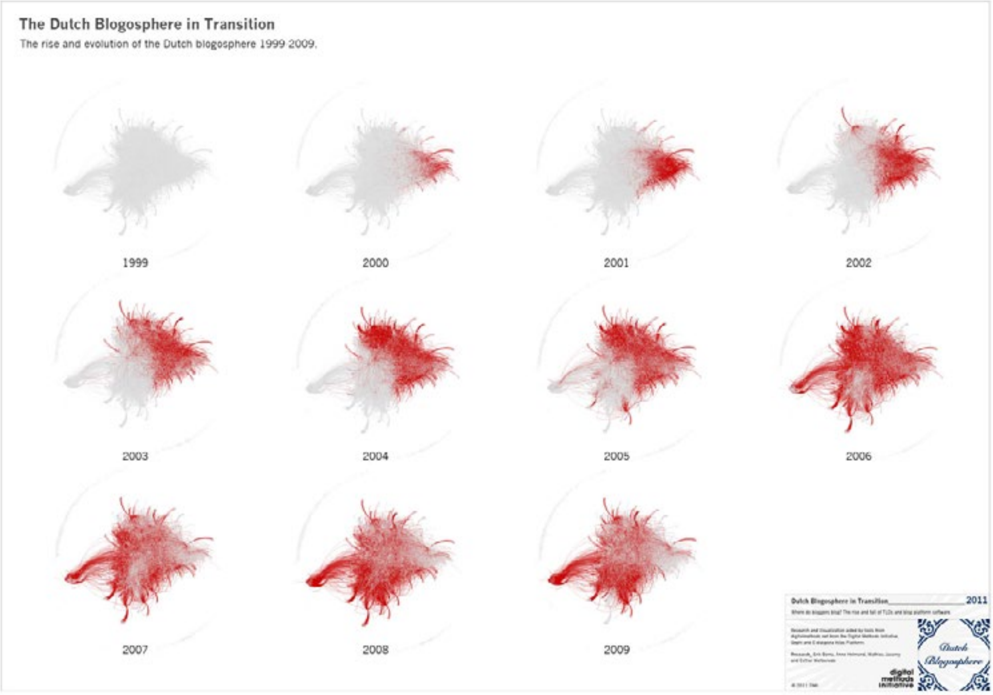


Figure 16: The Dutch blogosphere in transition: the rise and evolution of the Dutch blogosphere 1999-2009. Mapping the outlinks of the blogs retrieved from the Wayback Machine from 1999 until 2009 allowed to go back in time and study how and where the Dutch blogosphere originated. The network is made with the Internet Archive Wayback Machine Network per Year tool and networks per year are overlaid in Gephi. Thereafter a co-link analysis was applied and the graph was drawn using the Force Atlas 2 layout algorithm. The figure shows the rise, evolution and first signs of decline of the Dutch blogosphere. Grey depicts the hyperlink network of all years together and red the blogosphere of a particular year. The first Dutch bloggers starting mid 1999 were not interlinked into a 'sphere', so the beginning of a Dutch blogosphere can be traced back to 2000. Visualization created with Gephi and Adobe Illustrator, 2011. Source: Weltevrede and Helmond 2012a.

To identify nodes in the blogosphere as actors, they were redefined on a URL level. This required an additional analytical step because not all URLs followed the same pattern. In that way URLs pre-formatted analysis, although not all in the same standardized manner. With most websites 'actor' equals 'host' (such as example.com) while actors on blog software usually are defined before the host on a subdomain (such as example.blogger.com), actors on personal homepages are often defined by their ~ after the slash (such as xs4all.nl/~example), just like microbloggers on Twitter (such as twitter.com/example). In the following analysis actor definition these distinctive 'URL patterns' are formalized in the network.

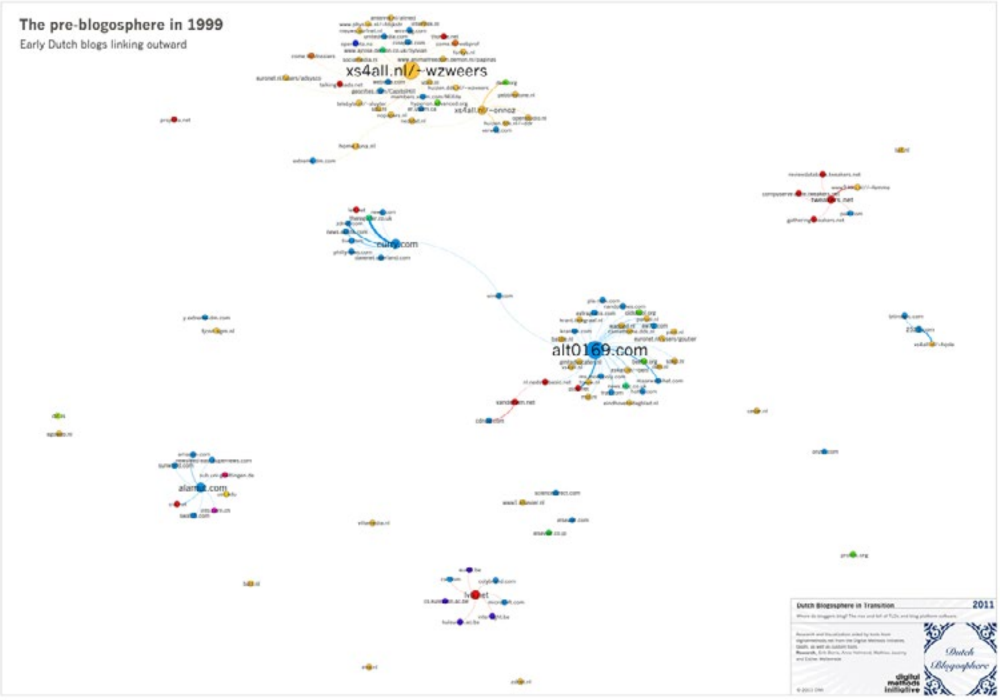

The pre-blogosphere in 1999
Early Dutch blogs linking outward
xs4all.nl/~wzweers
xs4all.nl/~onnoz
alt0169.com
alamut.com
Dutch Blogosphere in Transition
2011
Dutch Blogosphere
digital methods initiative

Figure 17: The pre-blogosphere in 1999: early blogs link outward. The network is made with the Internet Archive Wayback Machine Network per Year tool and visualized with Gephi using the Force Atlas 2 layout algorithm; colors were produced with the modularity algorithm. Note that in this figure no co-link analysis was performed. Some of the known Dutch bloggers, as mentioned in Meeuwsen (2010), together with less well–known bloggers, are present but do not form a blogosphere yet. Most notably Alt0169, ~wzweers and ~onnoz reach out to other Dutch blogs and may be seen as an effort to establish a community between blogs. Visualization created with Gephi and Adobe Illustrator, 2011. Source: *Weltevrede and Helmond 2012a.*

The evolution of the Dutch blogosphere

Mapping the outlinks of the blogs retrieved from the Internet Archive from 1999 until 2009 allowed to go back in time and study how and where the Dutch blogosphere originated. Using the fine-grained actor definition, the network was visualized with Gephi for each year. Figure 16 shows the rise, evolution and first signs of decline of the Dutch blogosphere, grey depicting the hyperlink network of all years together and red the blogosphere of a particular year. The first Dutch bloggers starting mid 1999 were not interlinked into a ‘sphere’, which leads to the conclusion that this visualization captures the beginning of the Dutch blogosphere in 2000.

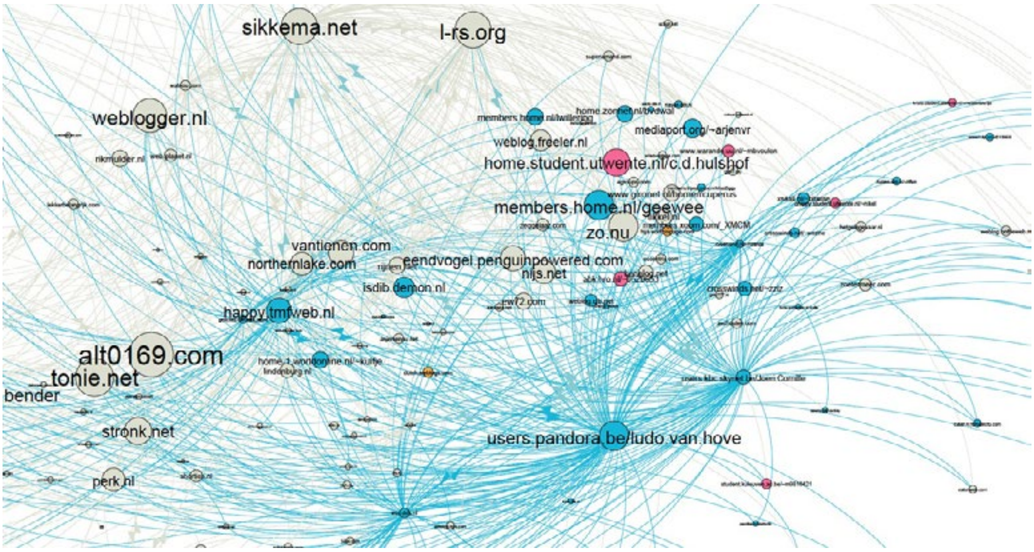


Figure 18: Partial screenshot of the reconstructed Dutch blogosphere in 2000. All blogs retrieved from the Wayback Machine have been color-coded based on the type of host. Blue nodes are personal homepages, pink nodes are student pages, and yellow pages are early blog platforms. Bloggers on personal homepage providers (blue) and student pages (pink) dominate. For the full interactive version with the G-Atlas tool see Weltevrede and Helmond 2011.

In 1999 the map (not displayed) only showed four nodes, not linking to each other but present because they received at least two links from the selected starting points. The four nodes were Nedstat, Nedstatbasic, Wired and a Dutch blog by Wessel Zweers, a.k.a. ~wzweers. A familiar node on the map is Wired, a technology magazine also prominent in the American early blogosphere. The only Dutch blogger on the map was hosted on one of the oldest Dutch hosting services providing free personal homepages, 'De Digitale Stad' (DDS, Digital City). Well–known Dutch blogs from that period, like Sikkema, Prolific and Alt0169 are notably absent because they did not receive two links from the starting list's blogs. Figure 17 shows that some of these well-known Dutch bloggers, as also mentioned in Meeuwsen (2010), together with less well-known bloggers, were present but did not form a blogosphere yet as the links were primarily one-directional. Most notably Alt0169, ~wzweers and ~onnoz reached out to other Dutch blogs and may be seen as an effort to establish a community between blogs. Exemplary are links to blogs listing blogs, like beboo.org/metalog, listing the top 50 (international) blogs.

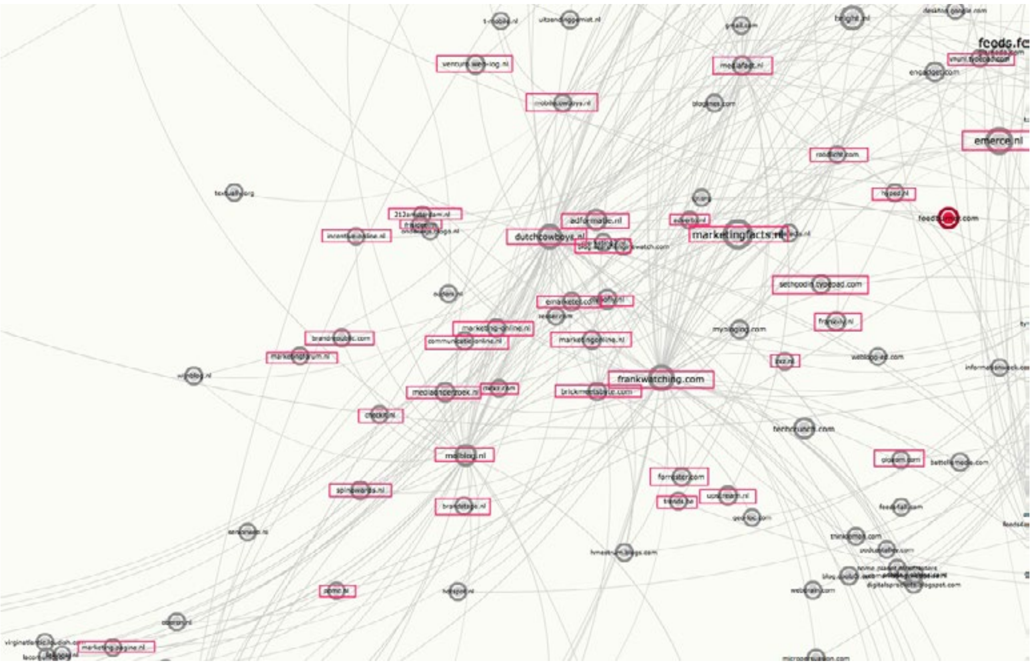


Figure 19: Partial screenshot of the reconstructed Dutch blogosphere of 2005: a cluster of marketing blogs (in pink). Marketing blogs are marked in pink. The Dutch marketing cluster emerged in 2005 and was still a very dominant cluster in the Dutch blogosphere in 2011. Data retrieved from the Wayback Machine. Visualization created with G-Atlas and Adobe Illustrator, 2011. Source: Weltevrede and Helmond 2011.

Cluster analysis over time

In the beginning, in 2000 the Dutch blogosphere was dominated by bloggers on personal homepage providers (blue) and student pages (pink) (see Figure 18). The right side of the blogosphere shows a cluster of Dutch homepages (~) and student homepages. The free homepage provider DDS and Dutch Internet service provider XS4ALL were the most prominent providers. The larger nodes in the center are the founding blogs of the Dutch blogosphere, such as Alt0169, Sikkema, S-lr, Smoel, Rikmulder, Tonie, Prolific, Pjoe, Stronk, Ben Bender, Vandenb, Retecool; they form a closely linked cluster. Alt0169.com, a heavy linker in 1999 but not receiving any links back, is a central node in 2000. Figure 19 shows the Dutch marketing cluster, which emerged in 2005 and still was a very dominant cluster in the Dutch blogosphere at the time of this empirical study. Another distinct cluster in the later blogosphere was the Blog.nl cluster. Blog.nl has a very distinct shape because all Blog.nl blogs listed and linked the other blogs on that platform as can be seen on the right in Figure 21.

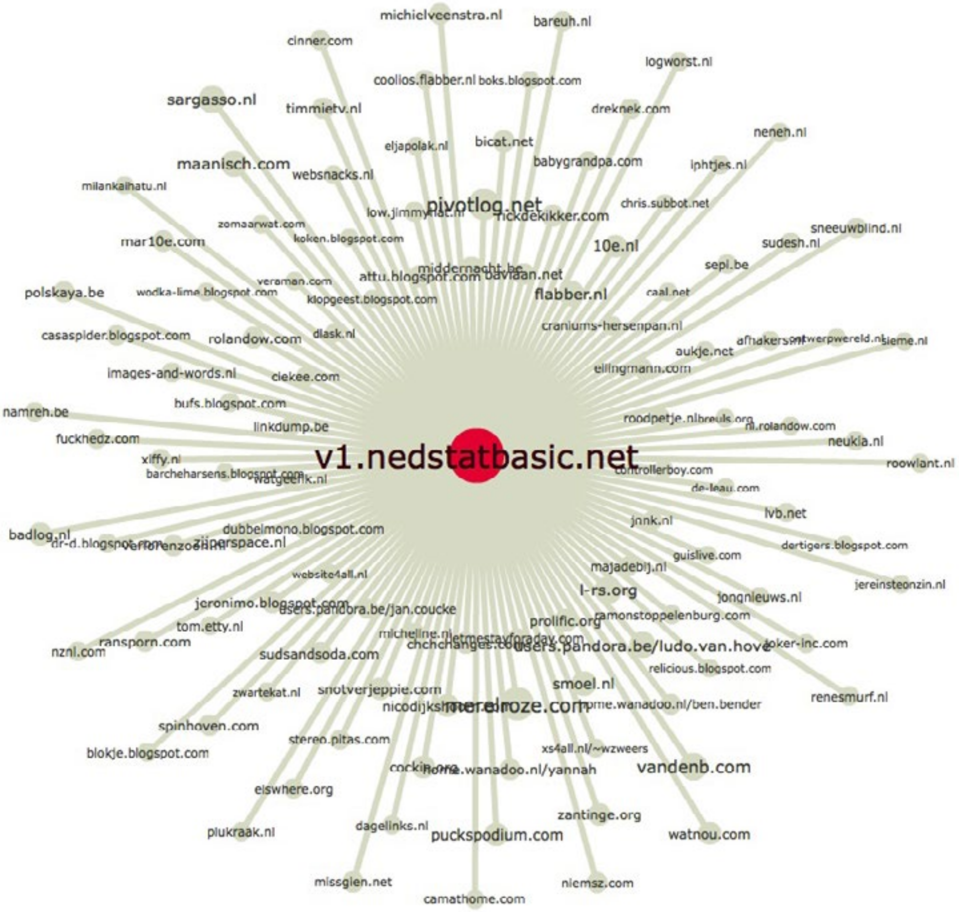


Figure 20: Partial screenshot of the reconstructed Dutch blogosphere in 2004: all co-linked Bloggers also link to the Dutch web statistics service Nedstat Basic. Data retrieved from the Wayback Machine. Visualization created with G-Atlas and Adobe Illustrator, 2011. Source: Weltevrede and Helmond 2011.

By means of the same method for coding blog platforms for the platform analysis several categories were created to trace specific transitions in the Dutch blogosphere by coding them in Google Refine: Homepages, University Homepages, Blog Related Services, Platforms, Social Media Platforms and Statistics. The emergent categorization was created by reading the URLs, and iteratively complemented with new findings throughout the project. This categorization enabled to color actors belonging to a specific category in Gephi, making it easier to locate actors and track changes over time. This method made it possible to look at the role of blog related services and social media in the blogosphere over time, which is reported on in the following.

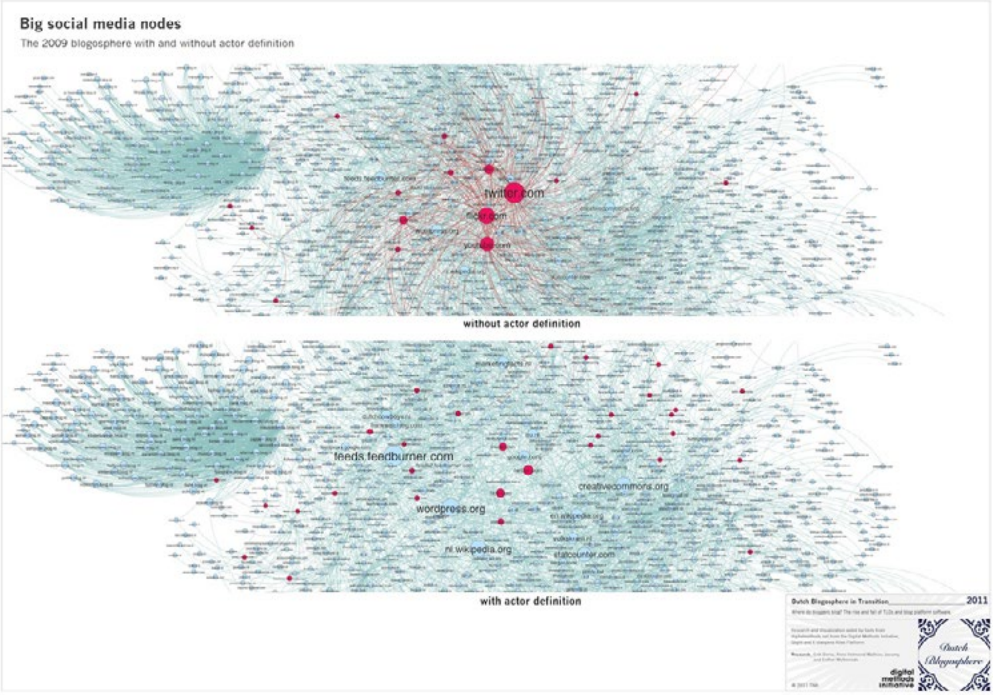


Figure 21: The Dutch Blogosphere in 2009: a comparison of different actor definitions. Social media platform nodes are highlighted in magenta. Both graphs show the full hyperlink network of the 2009 Blogosphere (i.e. without co-link analysis). Top: nodes represent host names. Bottom: nodes represent platform specific actor definitions such as user profiles. Through a more fine-grained link analysis (bottom) it becomes possible to analyze the role of social media platforms within the Dutch blogosphere in more detail. Data Retrieved from the Wayback Machine. Visualization created with Gephi and Adobe Illustrator, 2011. Source: Weltevrede and Helmond 2012a.

Blog related software: statistics

The newly defined blogosphere includes a variety of blog-related actors. The blogosphere does not only take shape by the interconnections between the blogs themselves, but also by the interconnections between the blogs and other actors, such as links to external (blog) services and to the blog software homepages. Blog-related services include portals, manual and automatic blog indexers, external comment services and statistics providers. One of the most prominent nodes since 1999 has been Nedstat. Nedstat—and its basic/free service Nedstatbasic—is a Dutch service providing statistics for web masters and bloggers about their visitors, present in the blogosphere together with other statistics providers. Most bloggers publish their statistics, which supports the claim that ‘the blogosphere is obsessed with measuring, counting,

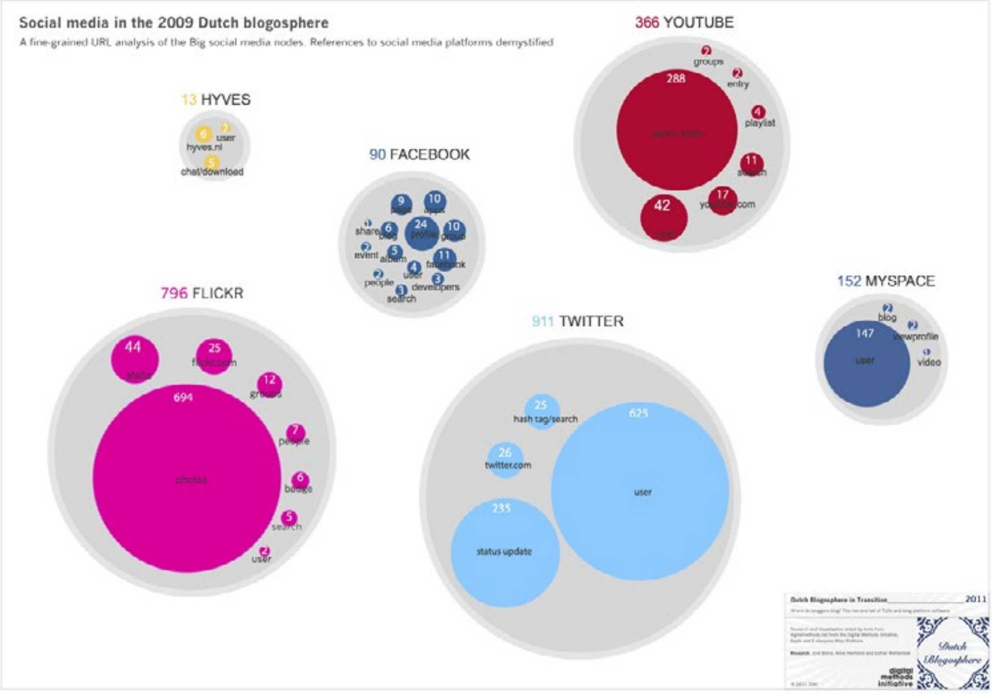


Figure 22: Different types of links to social media in the 2009 Dutch blogosphere. The grey circles represent social media platforms in the 2009 blogosphere. The colored circles represent different types of platform links such as user pages, hashtag queries, and status updates. Each circle is scaled proportionally to the number of links received for a particular actor definition. Comparing the various social media platforms, the results suggest that some platforms can be defined as 'media sharing' platforms, such as YouTube and Flickr, which mainly consist of embedded content links in blogs. Data retrieved from the Wayback Machine. Visualization created in Adobe Illustrator, 2011. Source: *Weltevrede and Helmond 2012a.*

and feeding' (Lovink 2007, n.d.). Zooming into the node (see Figure 20) all linked bloggers are shown, presumably using Nedstat as their statistics provider.

Social media analysis

The early blogosphere is characterized by larger nodes such as Alt0169, Sikkema, ~wzweers, the founding fathers of the Dutch blogosphere. The heydays of the Dutch blogosphere are characterized by the rise of specific clusters, such as the marketing cluster and the blog platform cluster of Blog.nl, and by the rise of blog-related services such as statistics. The later period is characterized by social media, the widgetized self and content links. In this social media research project methods are developed to analyze more closely the practices between blogs and social media.

Schaap (2004, 5) empirically researched what he calls 'the dichotomous nature of the Dutch blogosphere' caused by the clear divide between two distinct types of weblog forms: the linklog and the lifelog. The 'platformlog' was added as a third type of blog with particular characteristics. Whereas lifelogs primarily posted about their daily life in a diary style and in most cases only linked to their about page, their off-line contexts and other bloggers, the linklogs linked abundantly to other blogs and media when pointing out the best of the web (Schaap 2004). The platformlog embedded and linked content from social media platforms like Flickr, YouTube and Facebook and referred to the author's presence on these platforms in sidebar widgets. The platformlog was often used to present the widgetized self (Baym 2007), or the distributed self across social media platforms (Helmond 2012). Whereas in the mid and late 1990s the self was defined on the personal homepage and later on the blog, nowadays the self is also defined and performed on social networking sites and content platforms. Blog software made the creation of the widgetized self popular with its easy drag and drop widgets allowing bloggers to easily embed content from their other platforms into their blog via the sidebar. The sidebar is no longer only used to link to other bloggers, using the blogroll, but also to link to the self on other platforms such as Last.fm for music, Flickr for photos and YouTube for videos. As outlinks were collected from the front-page and these links were subsequently co-linked, the widgetized self in the sidebar on the front page could be captured as a new actor in the blogosphere. In traditional hyperlink analysis social media nodes were disproportionally large as all references were collapsed into one node.

Comparing the 2009 blogosphere *with* and *without* the custom actor definition (Figure 21), it became apparent that the social media platforms privileged a more fine-grained analysis. Social media are the big nodes in the network without actor definition; with actor definition, however, the social media platforms seem to reduce significantly in prominence in the blogosphere demonstrating how digital methods have to respond to the volatility and evolution of digital media and their cultures of use.

The question then arises what do people link to in social media: to user pages or to content (such as video, photo, status update)? Figure 22 shows the large social media platform nodes, containing smaller nodes. Comparing the various social media platforms, the results suggest that some platforms can be defined as 'media sharing' platforms, such as YouTube and Flickr, which mainly consist of embedded content links in blogs. In the blogosphere map with actor definition, these nodes decrease in size. Facebook is a relatively small node in the Dutch blogosphere and the links it receives dissolve into a diverse set of profiles, pages, apps, events, and groups. Hyves—the late Dutch social network—is one of the smallest social media references. Although the Dutch blogosphere privileges Dutch software and platforms, this is not reflected in social media platform links. Twitter, the largest node in the network, is a platform mainly receiving links to user pages. This means that bloggers refer to themselves or to friends on the micro-blogging platform.

Link analysis needs to evolve and adapt to analyze the share of social media in blogosphere networks. The results of this study suggest that the uniform large platform nodes are misleading. One of its main findings is that link analysis zooms out to look at platforms as a whole and treats the entire platform domain as the node; the individual content link and the individual author thus disappear. Instead, the platform nodes require a more nuanced exploration.

Conclusion

In digital research the volatility of digital methods signals the intricate relation between medium and method. The research aim is brought in dialog with the device-driven research affordances and their cultures of use. Digital research is thus always in part medium research, because it draws attention to the specific purposes and use practices inscribed into digital devices and the consequences of specific configurations on the findings. For example, the study in this chapter empirically found that co-link analysis, traditionally used for 'issue research', turned into 'infrastructure research' when brought in configuration with the specific affordances of the linking practices of the blogosphere, as well as how the blogs are archived and made available through the Wayback Machine.

The study also captures the volatility of digital culture more generally, as it captures and visualizes the transitions and developments in the Dutch blogosphere from open-ended DIY technologies to off-the-shelf (social media) platforms. In the following chapter the repercussions of online update cultures for device-driven research are explored more in detail. This chapter thus also contributes to the body of literature on the blogosphere and social media platforms by proposing new methods to empirically investigate transitions in the historical blogosphere over time and by mapping the rise of social media platforms from within. Hence a method was developed and described to create a so-called structural blogosphere on the basis of the medium-specific characteristics of the Internet Archive's Wayback Marchine, allowing for the re-construction of a blogosphere on domain level and not on post level. The advantage of this method is that it allows for a 'structural' blogosphere analysis instead of an 'issue' or 'event' analysis. A structural analysis enables software and platform analysis; in this case it served as a new way to study the nationality of blogs. The study contributes to researching the evolution of blogging technologies by looking into TLD, platform and software usage. The study suggests that Dutch bloggers increasingly blogged on in the .nl space despite the more general trend of software concentration and domination of actors like Blogger and WordPress.

This chapter further develops three analytical techniques and methods to study the national historical blogosphere: URL, source code and hyperlink analysis. URLs are very rich sources of information, often following a certain syntax, which makes them very suitable for analysis. In this study URL analysis was performed in two ways: TLD and platform analysis. The introduced source code analysis contributes to the study of software in general and, more specifically, to the study of national software. The method developed provides insight into the software powering a blogosphere. Further research may include a fine-grained feature analysis over time, especially emphasizing collaborative and discursive features such as the comment, plugins and the permalink. The contribution to link analysis considers ways to treat the large platform nodes in network visualizations by introducing two techniques, the first being an actor definition and the second a fine-grained social media analysis. Traditionally the host is considered as the actor, but when dealing with platforms the actor, or blogger in this case, is often defined after the slash. By detecting URL patterns, new actor definitions may be implemented *before* co–link analysis. Fine-grained social media analysis is similar in technique, but instead of only looking for actors, it tries to distinguish actor and content links. The analysis is performed *after* co-link analysis.

The evolution of technologies in the Dutch blogosphere may also be approached in terms of the transitions in the ‘grammars of action’ encoded into them (Agre 1994). A shift is noticeable where increasingly messy, general-purpose blogging features and functions evolved into closed and off-the-shelve software, rendering them more easy useable but at the same time, by introducing more fine-grained and scripted grammars, more difficult to tinker with (see also Van Dijck 2013; Zittrain 2008, 104–107). These features and functions as affordances imply that they increasingly pre-structure potential actions that can be assembled like building blocks by third parties, including digital researchers. The following chapter investigates the implications for device-driven digital research of volatile media and their evolving research affordances.

Google algorithm changes and the volatility of method

Chapter 5

IN THE PREVIOUS CHAPTER the volatility of digital methods was introduced by focusing on the intricate relation between medium and method. In this chapter, I continue this inquiry by focusing on one key digital medium which is in a state of perpetual beta, namely Google Web Search. In device-driven research digital media are being operationalized to operate as devices in the method pursuing myriad research purposes. This chapter focuses on the challenges and opportunities of a volatile medium such as Google Web Search for digital research. In so doing, the chapter also reflects on the distinction between medium research and social research in the context of the various configurations of the Google Web Search engine for the purpose of digital research. In their 1998 article, Google founders Sergey Brin and Larry Page introduced the then-current anatomy of Google, and stated that 'in addition to being a high quality search engine, Google is a research tool [...] for a wide range of applications' (Brin and Page 1998, n.d.). To what extent, how, and for what may the Google search engine then be used for digital research?

It is obvious that search engines are key devices in digital culture and social life as they play a crucial role in selecting what information is considered most relevant. Search engines allow users to search and navigate massive databases of information, map user preferences against each other and suggest information to the user. Recommendation algorithms, such as the ones employed by search engines to output their results, are designed to calculate what is 'relevant' and in doing so make a suggestion from a large pool of alternatives. There have been many critiques on search engines' notions of relevance. Inquiries into the definition of 'relevance,' however, have struggled with the fact that an objective baseline for what constitutes relevant results lacks, as it not an inherent trait of the results, but instead always in response to an informational need (Rieder 2012). Research has shown that Google employees generally equal 'relevance' to customer satisfaction as well as to some notion of accuracy (Van Couvering 2007; Gillespie 2011). Early search engine research furthermore looked at the biases of search engines and noted tendencies toward already popular, English-speaking content and commercial content (Introna and Nissenbaum 2000; Halavais 2008; Granka 2010). Legal scholars debated the supposed neutrality of search engine results (Grimmelmann 2009; Pasquale and Bracha 2007) and considered search engine recommendations as acts of freedom of expression (Van Hoboken 2012). The transparency and bias in algorithms was several times approached from a legal point of view, and these issues are increasingly investigated from a non-legal point of view, which focus on normative concerns involving algorithmic transparency and bias (Sandvig et al. 2014; Barocas, Hood, and Ziewitz 2013; Gillespie 2014).

This chapter further develops the device-driven approach and investigates how search engines may be repurposed as devices to do research with, which was previously referred to as 'search as research' (Rogers 2013b), or as what I once termed 'using Google as a research machine' (Weltevrede 2010). Since search engines, and Google in particular, are commonly considered to be the starting point to the web, recommendation algorithms may be considered as dominant 'knowledge logics' as they provide a means 'to know what there is to know and how to know it' (Gillespie 2014). This chapter inquires into the affordances of Google Web Search for specific modes of digital research that make use of the engine's knowledge logics. While at first sight such an approach may be critiqued for using corporately owned algorithms, it provides an opportunity to discuss some of the limitations and opportunities of digital methods and their resemblance to many of the issues that trouble digital social life much more in general, such as medium dependency, the volatility of methods and the black-boxedness of knowledge technologies. Especially the latter two points are pursued further in this chapter. The epistemic difficulty explored by digital methods is not just a problem of cultural and social studies but rather one of many social practices involving the collection, management and analysis of digital social data, as was also discussed in Chapter 1 and Chapter 2. In this chapter in particular I will empirically explore repurposing Google, to highlight the digital device as an object to study digital culture and social life, which it is supposed to enable.

The emerging field of software studies has begun to explore how computational objects such as recommendation algorithms capture social and cultural life, as well as how they format and push into it. Algorithms can be seen as a specific type of computational object which may be distinguished from the program or code for example, as it focuses on the strategy, intent or plan embedded in the software. In this chapter I build on the strand of software studies that seeks to study the operational principles of algorithms without necessarily extrapolating findings from source code (Rieder 2013; Bucher 2012a; Rieder and Röhle 2012; Feuz, Fuller, and Stalder 2011; Gerlitz and Helmond 2013; Sandvig et al. 2014). Such approaches are highly informative for digital research as they help to identify the methods embedded in algorithmic media. By exploring the notion of the black-box and possible ways of studying its workings via software studies, I develop an approach that combines different types of methodological and conceptual resources on how Google's algorithms can be repurposed for research.

This chapter can be read as a contribution to the rich and growing field of algorithm studies, such as the auditing algorithm approach, as it focuses on the technical logic of the Google search engine to the extent to which it is productive for digital research. Not only has there been an epistemic shift in how Google changed our idea of what 'relevant' information is, Google's idea of relevance is also subject to change, as is what can be learned when using the engine for research. In an algorithm study over time I discuss the modes of research enabled by Google when regarding it as a research device.

I will explore various modes of research questions that can be asked by using Google as a 'research machine' and will pay attention to how the search engine can be configured for social research, even though the precise inner workings of Google may remain obscure. In concurrence with recent software studies work, the detailed examination of various sources which discuss Google's algorithms, such as whitepapers, patents, trade press, and help documentation, allow to recognize how Google captures, formats, and recommends data and how this knowledge about the operations and effects of algorithms can be taken into account when pursuing digital research with the engine. By inspecting key algorithm changes throughout the years, this chapter inquires into how these updates not only affect but also afford different modes of digital research that use Google as a source for data. The modes of research discussed include both social research and medium research projects. These modes of research are illustrated with seminal case-based projects. It is thus shown that various settings and configurations can make Google into a multifarious, albeit volatile research tool for various research purposes, as algorithm changes can enable specific research modes yet make others obsolete. The exploration of the volatility of the continuous update culture of algorithmic media will thus inform the volatility of research with digital research methods.

This chapter deals therefore not exclusively with algorithmic media as technology or technique. Instead I seek to bridge the distinction between the epistemic and the technical value of knowledge by readjusting the research objective to take up the challenge of how to research *with* algorithmic recommendation to make findings about what can be known with, and about, the algorithmic medium. Through the inquiry into the digital cultures, techniques, and algorithms that order and recommend information this chapter is consequently also a contribution to 'web epistemology' (Rogers 2004). By studying the medium as a techno-epistemological device, it thus becomes possible to research with the medium. In other words, I approach the algorithmic medium as containing the conditions of possibility for what can be known with it and how. Through the notion of research affordances I offer an empirical exploration of how algorithmic media such as Google Web Search activate ways of doing, knowing and operating, and the implications that can be drawn from them. While this chapter focuses on Google to illustrate how research that takes web method and medium-specificity seriously may be informed, the principles put forward here can be used to repurpose other digital devices as well and recur throughout the case studies in the other chapters.

Studying black-boxed algorithms

A simple search bar invites users to get access to 'the world's information' by entering a query into Google's Web search (Google n.d.). This search box, however, hides many algorithms processing the query and recommending the most relevant results. Google's algorithms are proprietary and secret, and hence are unavailable for public scrutiny (Pasquale 2009). Popular scientific literature often describes such algorithms as black-boxes. Historically, the notion of the black-box refers to the box hiding radar equipment in WWII. Confronted with an enemy black-box engineers had to deduce its functioning by observing the relationship between input and output (Ashby 1956; Wiener 1961; Von Hilgers 2011). When considering Google's algorithms, a user can only enter input in the form of a search query and receive output in the form of result listings, but what remains hidden is how the recommendation engine works; although certain details are known, others, such as how specific criteria are measured, weighed against each other, and which criteria override one another remain obfuscated.

In a second meaning of the term, STS scholars Trevor Pinch and Wiebe Bijker (1984), as well as Bruno Latour (1988), used the notion of the black-box to refer to what they call a stabilized and closed object, that is, when the messiness of the object's making, as well as the ensemble of associations that went into it, are forgotten or are no longer relevant to its users. In this sense Google may also be considered a black-box, as it is the main entry point to the web for users often having little interest in how specific results are returned in response to a query, especially since the basic search bar has remained relatively unchanged throughout the search engine's existence.[53]

There are a number of approaches to study algorithms. For instance, the academic field Critical Code Studies (CCS) focuses on the cultural significance of computer code, where reading the code is one of the key approaches. Mark Marino, who advanced CCS in 2006, states that it is a method to 'read and explicate code the way we might explicate a work of literature' (Marino 2006, n.d.). In addition to not being available for public scrutiny, there are some notable limitations to what can be known about algorithmic media when resorting to the reading the code approach. Wendy Chun argues that there is a tendency among scholars to appreciate source code as the truth or essence of software, thereby turning source code into a fetish. She warns that source code should be treated as a re-source instead of the cause for computation as 'source code only becomes source after the fact'; only when the algorithm has been executed can its outputs be deduced from the code (Chun 2011). Another issue with (just) studying code is addressed by Lev Manovich when he points out that the

[53] See this video produced by Google showing the design of the frontpage over time (Yehoshua and Nath 2015).

‘reading the code’ approach creates the illusion of a static and definite text that can be studied, ‘but this is an illusion, and we have to accept the fundamental variability of the actual “software performance”’ (Manovich 2012a). Instead of dealing with static text, software is dynamic and always in dialogue with the user through their interaction with the software. As Manovich elaborates: ‘what the user experiences as a single web page may involve continuous interactions between dozens or even hundreds of separate software processes’ (Manovich 2012a). Others also noted that even a detailed formal description cannot fully document a program’s inner workings (Goffey 2008; Rieder and Röhle 2012; Bucher 2012a) as the source code alone does also not suffice to understand the implications of algorithms; reading code is not able to show all specific outcomes of the algorithms—especially when algorithms are highly iterative or when result pages are an aggregate of a large number of individual calculations.[54] In fact, the algorithms of digital media such as Google are ‘likely so dynamic that a snapshot of them would give us little chance of assessing their biases’ (Pasquale 2009).

Consider how Andrew Goffey stresses the ‘pragmatic dimension of algorithms’ and explains that ‘algorithms do things, and their syntax embodies a command structure to enable this to happen’ (2008, 16–17). While in computer science (applied mathematics) the algorithm is first and foremost a step-by-step method prescribed by given parameters by which a task is accomplished, Goffey draws attention to how algorithms are more than their formal expression and opens up ways to study the social and cultural role of algorithms beyond the source code: ‘algorithms act, but they do so as part of an ill-defined network of actions upon actions, part of a complex of power-knowledge relations, in which unintended consequences, like the side effects of a program’s behavior, can become critically important’ (2008, 19). Similarly, Wendy Chun points out that an algorithm is a ‘strategy, or a plan of action—based on interactions with unfolding events’ (2011, 126). Google’s algorithms, too, are embedded into different types of digital practices producing the data which the algorithms process. By drawing attention to the pragmatic functioning of algorithms it becomes clear that, if we want to use recommendation algorithms for social research, we should shift focus to the algorithms in action. More recently, approaches to study algorithms are being developed that do not necessarily focus on the source code but instead are interested in the specific strategies or intents embedded in algorithms, which will be discussed shortly. I contribute to this by arguing that it is instructive to look at the potential of algorithms as techno-epistemological devices, by repurposing its analytical affordances and to make the algorithmic medium part of the operationalization of the research question.

[54] That digital black-boxes are highly iterative makes them different from the traditional notion of a black-box where the internal workings would not be susceptible to continuous change.

One way to do this stems from the original notion of deducing the inner workings of a black-box. Reverse engineering is a diagnostic approach that observes the relationship between input and output and thus focuses more on algorithms in practice. The device-driven research approach put forward here has similarities with the reverse engineering approach put forward by software studies scholars Taina Bucher (2012) and Robert Gehl (2014). Both draw on various materials to research the ways in which software shapes digital culture and social life online. I am specifically drawn to Bucher's suggestion to develop methodological ways let the software 'speak' (2012a, 74). Following Latour's suggestion to invent tricks to let the non-human speak, she notes that making software speak as to read it in the way 'one reads any text' (Galloway 2004, 20) is not without its problems, most notably due to its black-boxed nature. Bucher therefore suggest to focus on 'the interface and the descriptions of its workings contained in external textual sources' (2012a, 74). The 'auto-ethnographic' approach she develops is highly individualized and personalized as the observation of the interface is confined to the 'me-centric' view of the researcher's own account (2012a, 81).

A second recent approach I want to mention here proposes to study black-boxed algorithms through a mode of research called 'auditing algorithms', which puts forward various methods to study the extent to which algorithms are discriminatory (Sandvig et al. 2014). The research design outlines a number of different possible ways algorithms might be brought in dialog with the research objective of the social scientific auditing method. Although the paper is a research proposal and the methods are not brought in dialog with the differences between various algorithmic media, the paper does flag this as an important aspect in the method and I thus consider algorithm audit a mode of device-driven research. Also the auto-ethnographic work of software studies scholar Taina Bucher in researching Facebook's Newsfeed can be considered a specific mode of device-driven research. One that, as opposed to the algorithm audit proposal, makes full use of the personalizing capacities of the Facebook Newsfeed algorithm. I am particularly interested in how these algorithmic devices intervene with the method and how digital researchers can respond to the operative capacities of algorithmic media. I do so by inquiring into the configuration of the device and the research apparatus.

As discussed before, digital algorithms are iterative, continuously changing and are often aggregates of calculations and as such their exact workings at any given time are hard to retrieve. Although we might not know the exact details about how the recommendation algorithms weighs different factors, other important facets are available. Most importantly, we know that Google acts on various data points in web pages, videos and certain social network sites, as well as the way in which it captures and formats them, and many more. With digital media it is key to recognize which data points are captured, how they are used by the medium, for which purposes, and to explore the research affordances for digital research. Consider for example how the various ways in which Google localizes otherwise 'placeless' web content allows researchers to inquire into the specific national or local information cultures the search engine creates.

The device-driven approach combines different types of methodological and conceptual resources to study which data points Google captures, and how it acts on them, in order to gauge the opportunities and affordances of Google as a techno-epistemic device. The main sources are patents, which are useful because they give insights into the intentions behind algorithm updates, blogposts about updates to confirm changes, in combination with empirical device-driven projects. Such an approach offers a useful method to start understanding how algorithms capture and format social life. I am therefore specifically interested in the pragmatic dimension of algorithms and study, and periodize, the ways in which the Google algorithms change. Following the technicity of how algorithms (pragmatically) capture, format, and recommend prescribes 'what there is to know' and how to know it, I group algorithm changes and couple them with modes of research. This approach seeks to understand what Google says about its algorithms in relation to what its search algorithms do. Hence I shift the research objective from knowing algorithms to researching *with* algorithms and inquiring into how the algorithm intervenes in the method. With it, the main focus in digital research moves from studying source code or interfaces to the device-driven perspective.

In the next sections I will first focus on the canonical PageRank algorithm through Page and Brin's whitepaper (1998), key patents and empirical projects. I will then use a variety of sources to study algorithm changes as well as their different research affordances (as Google Web Search's current algorithm is not only PageRank but consists of over 200 signals and metrics).

PageRank and the neutralizing engine

Since its introduction in 1998, Google quickly became the dominant engine in a growing number of countries. In 2012 Google had actually become the leading search engine across the globe, with the exception of a few countries.[55] As the leading engine that indexes and organizes what has been termed the 'destination web' (Berry 2011a) or 'static web' (Searls 2005), it is one of the recurring devices used in digital methods

[55] These countries include China with Baidu, Japan, Hong Kong with Yahoo!, Russia with Yandex, Czech Republic with Seznam and South Korea with Naver (Landry 2012).

research, and for different analytical purposes. While technically speaking the algorithm may be volatile, ethically and politically the mediating processes are stabilized to form part of a socio-technical environment enabling the dependencies on algorithmic media to go unnoticed, and thus often being perceived as a 'neutral' intermediary.

This section treats the founding PageRank algorithm and discusses the modes of digital research it enables as well as how specific choices in the algorithm affect research. In this section I focus on how with PageRank the search engine has been semantically and rhetorically positioned by its makers and has allowed it to operate as a 'neutral technology'. The careful articulation of the algorithm as automated and impartial seeks to certify its relevance as credible, positioning itself as a reliable socio-technical actor, and to maintain its apparent neutrality, considering the many evaluations it makes. This articulating and positioning of the algorithm is crucial for its function as dominant knowledge logic. By dissecting how Google positioned the PageRank algorithm as a neutralizing device I introduce what we can know through it and consequently introduce three types of empirical modes of research performed with Google results.

The 'neutrality' of Google as a search engine is first of all established since it defers some of the responsibility to the digital practices it indexes and uses as a key metric in the calculation of relevance (Brin and Page 1998). This registration and metrification of practice by the Google algorithm may be gauged by turning to PageRank's roots. Although search engines emerged in the field of information retrieval, an often-overlooked methodological connection is that PageRank has its roots in social scientific research (Marres 2012). Bernhard Rieder (Rieder 2012) connects PageRank more specifically to concepts and methods developed in sociometry, citation analysis, and social exchange theory by studying the references used in the two founding articles (Brin and Page 1998; Page et al. 1999) and the two most important PageRank patents, 'Method for node ranking in a linked database' (Page 2001) and 'Method for scoring documents in a linked database' (Page 2004). Especially the patents are very rich in references to graph theory and citation analysis within social sciences of the 50s and 60s.

Page viewed the web as a social system by scoring pages based on the number and importance of links pointing to them using the 'collective intelligence' of the web. Whereas in previous leading search engines such as Altavista, relevance was mainly determined by the number of keywords on a page matched a given query, Google abandoned the idea that the web was a 'flat corpus' and calculated the importance of every web document, independent of the page's content. 'A page has a high rank if the sum of the ranks of its backlinks is high' (Page et al. 1999). A link from a highly ranked site weighs more than a link from a site with a low rank. PageRank uses information external to web pages themselves—their backlinks—which provide a kind of recommendation and relevance. Additionally, it is recursive: links from important pages are more significant (Page et al. 1999).

PageRank assumes a model of user behavior with a built-in model for a 'random surfer', 'who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank' (Brin and Page 1998). They furthermore found that PageRank could be seen as a proxy for traffic (Brin and Page 1998). The Google PageRank algorithm may thus be seen to repurpose social scientific method for algorithmic recommendation and in doing so deflecting responsibility to the anonymous 'collective intelligence' of the web. PageRank gives web pages a higher score on the basis of what others find relevant (as expressed through the associations made by hyperlinks).

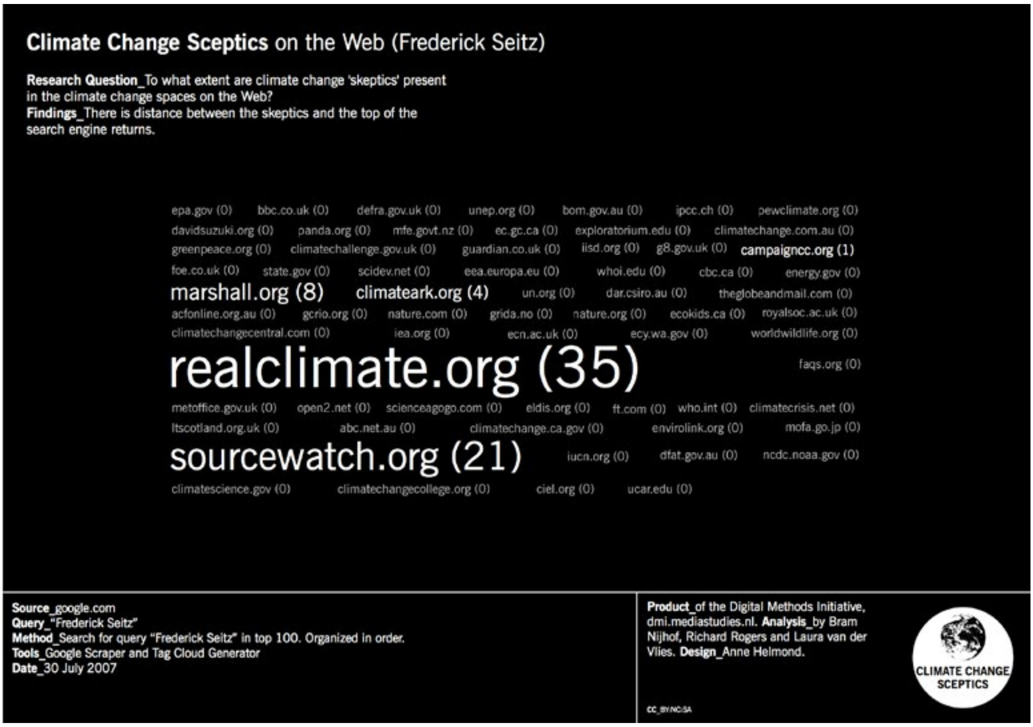
Let me give an example of the value of turning to the specifics of the social scientific logics built into PageRank. When investigating one specific parameter in the PageRank algorithm—the 'damping' factor α—Rieder concludes that accounts that interpret PageRank as a 'gigantic popularity contest' simply miss the target, because rather than showing us the popular, the way Google configures this parameter, 'it shows us the authoritative, or, to connect back to citation analysis, the canonical' (Rieder 2012). Rieder then argues that in that sense PageRank's vision of society is rather conservative and seeks to 'pick out the real leaders' (Katz 1953) and distributes visibility to them (Rieder 2012). The classic information retrieval notion of 'relevance', which is always conceived in relation to a specific informational need and not as an intrinsic value, is thus complemented by the sociometric notion of *capturing* 'status' and 'authority' (Rieder 2012). Interpreting the results of the pure PageRank is therefore an inquiry into the most authoritative sources for the query. However, this can also be seen as a circular process, as PageRank itself attributes authority by ranking sources highly and can thus be seen as a 'status-authoring device', one which *formats* social and cultural life (Rogers 2013b). It has been argued that the status of highly ranked sources is enforced by PageRank, since research has shown that users only look at the first few dozens of results (Hindman 2008; Jansen and Pooch 2001; Nanji 2014). With PageRank Google authors status on a global scale; repurposing this effect allows us to do research *with* that knowledge logic, to which I return later in this section.

In addition to deferring responsibility, a second argument often used by Google is that its selection and ranking are automated, thereby delegating it to the cold objectivity of the machine replacing the messiness of the human-edited directory. PageRank was developed in response to a major challenge in the search engine industry specifically, and arguably the web more in general. With PageRank, Google found a scalable solution to deal with the quickly growing amount of information on the web, making previous solutions like human-maintained indices obsolete, such as the then popular Yahoo! Directory[56], as well as the simple parsing of user-defined meta-fields or matching queries to words in the page, like the then popular Altavista search engine did. In addition to being more effective in dealing with the growing body of information on the web, the algorithmic automation of asking other sites for their recommendations allowed the search engine company to deflect any responsibility for the results. This is illustrated succinctly with case study analyses by James Grimmelman in "the Google dilemma" (Grimmelmann 2009), where he for instance analyzed how Google positions itself discursively in an advertisement placed next to racist results for queries like [jew] instead of intervening directly in the 'organic' results. However, while Google positioned the 'objective' algorithmic status authoring as more neutral than the subjective editorial practices of earlier media, Henk van Ess discovered in 2005 that at least some evaluations of results where performed manually via Google's eval.google.com. Google's spokesperson GoogleGuy responded directly to Henk van Ess' blog posting, stressing the relevance for Google at that time to discursively position the search engine as being fully algorithmic (Van Ess 2005)'.

In addition to deferring responsibility to the collective intelligence reduced into the hyperlink structure and automatization, the third way PageRank fosters neutrality is by the unobtrusive greyness of algorithmic media. 'Grey media' is described by Matt Fuller and Andrew Goffey (2012) as the 'frictionless' activity that is facilitated when media, practices, protocols and procedures get entangled and create black-boxes in the sense of accepted, more or less stabilized artifacts. This greyness occurs because there is also little interest in the finer details of the precise workings of the algorithm outside a few—but growing number of—areas in the field of computer science, social sciences and media studies, the professional field of search engine optimization (SEO), legislation and policy. Moreover, algorithmic media work best when they are perceived as intermediary and when their workings are not taken into account.

The discussion of how PageRank has been positioned as a neutral device (because it is objective, automated, and reuses the associations of others), allows us to recognize how Google's search results became a dominant knowledge logic (a status-authoring device) ordering the world's information on a global scale. (As we shall see later on, until 2000, Google had one ordering of results for a query globally.) This knowledge now allows me to elaborate on the first few modes of research that can (or could) be performed by using Google outputs. In other words, PageRank suggests a specific set of research questions which for instance are interested in which sources have most status for a specific query. How many of them are mainstream (compared with other

[56] Yahoo! discontinued the directory too (Bright 2014).



Climate Change Sceptics on the Web (Frederick Seitz)

Research Question_To what extent are climate change 'skeptics' present in the climate change spaces on the Web?
Findings_There is distance between the skeptics and the top of the search engine returns.

epa.gov (0) bbc.co.uk (0) defra.gov.uk (0) unep.org (0) bom.gov.au (0) ipcc.ch (0) pewclimate.org (0)
davidsuzuki.org (0) panda.org (0) mfe.govt.nz (0) ec.gc.ca (0) exploratorium.edu (0) climatechange.com.au (0)
greenpeace.org (0) climatechallenge.gov.uk (0) guardian.co.uk (0) iisd.org (0) g8.gov.uk (0) campaigncc.org (1)
foe.co.uk (0) state.gov (0) scidev.net (0) eea.europa.eu (0) whoi.edu (0) cbc.ca (0) energy.gov (0)
marshall.org (8) climateark.org (4) un.org (0) dar.csiro.au (0) theglobeandmail.com (0)
acfonline.org.au (0) gcrio.org (0) nature.com (0) grida.no (0) nature.org (0) ecokids.ca (0) royalsoc.ac.uk (0)
climatechangecentral.com (0) iea.org (0) ecn.ac.uk (0) ecy.wa.gov (0) worldwildlife.org (0)
realclimate.org (35) faqs.org (0)
metoffice.gov.uk (0) open2.net (0) scienceagogo.com (0) eldis.org (0) ft.com (0) who.int (0) climatecrisis.net (0)
ltscotland.org.uk (0) abc.net.au (0) climatechange.ca.gov (0) envirolink.org (0) mofa.go.jp (0)
sourcewatch.org (21) iucn.org (0) dfat.gov.au (0) ncdc.noaa.gov (0)
climatescience.gov (0) climatechangecollege.org (0) ciel.org (0) ucar.edu (0)

Source_google.com
Query_"Frederick Seitz"
Method_Search for query "Frederick Seitz" in top 100. Organized in order.
Tools_Google Scraper and Tag Cloud Generator
Date_30 July 2007

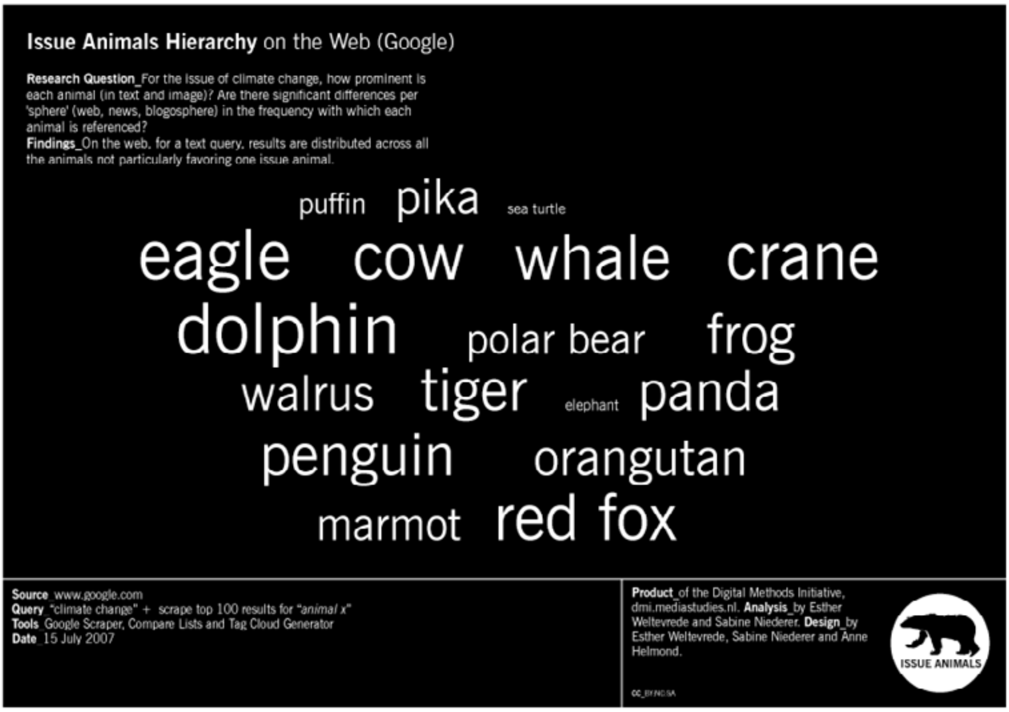
Product_of the Digital Methods Initiative, dmi.mediastudies.nl. **Analysis**_by Bram Nijhof, Richard Rogers and Laura van der Vlies. **Design**_Anne Helmond.

CC_BY.NC.SA

Figure 23: Climate Change Sceptics on the Web (Frederick Seitz). Tag cloud displaying the number of web pages mentioning skeptic "Frederick Seitz" in the top hundred unique hosts returned by Google for the query ["Climate Change"]. The order of Google results is retained in the visualization. It can be seen that Seitz is not well recognized by the 'top of the web'. Visualization created in Adobe Illustrator, 30 July 2007. Source: *Digital Methods Initiative 2009*.

media) and how many of them are alternative? How far from the top are specific sources ranked?

One of the Digital Methods program earliest empirical critical works using the outputs of Google search results studied the effects of PageRank by questioning to what extent the results returned provide mainstream or alternative voices in the top engine results (also see the work of Muddiman 2013; Eklöf and Mager 2013). Research with PageRank is for instance suitable for 'source distance' research, a technique developed during the first Digital Methods Summer School in 2007, which looks at the sources in which certain terms appear and how they rank for a certain query. The classic example project is 'climate change skeptics' which looks at the position of known climate change skeptics in the top results for the query ["climate change"] (see



Issue Animals Hierarchy on the Web (Google)

Research Question_For the issue of climate change, how prominent is each animal (in text and image)? Are there significant differences per 'sphere' (web, news, blogosphere) in the frequency with which each animal is referenced?
Findings_On the web. for a text query. results are distributed across all the animals not particularly favoring one issue animal.

Source_www.google.com
Query_"climate change" + scrape top 100 results for "*animal x*"
Tools_Google Scraper, Compare Lists and Tag Cloud Generator
Date_15 July 2007

Product_of the Digital Methods Initiative, dmi.mediastudies.nl. **Analysis**_by Esther Weltevrede and Sabine Niederer. **Design**_by Esther Weltevrede, Sabine Niederer and Anne Helmond.

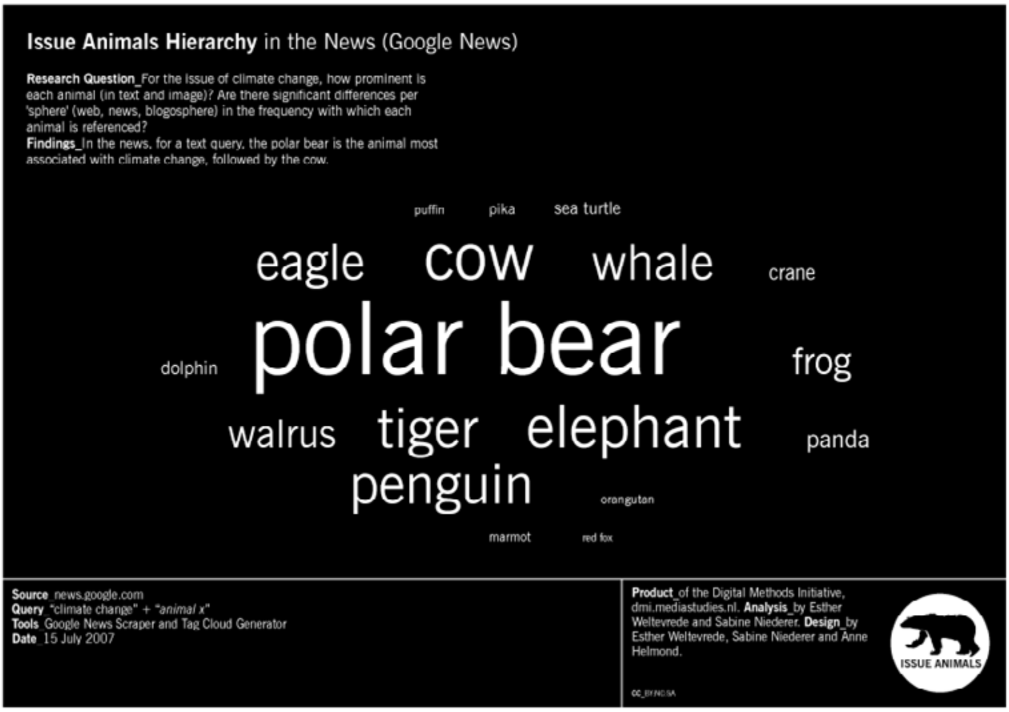
cc_BY.NC.SA

Figure 24: Issue Animals Hierarchy on The Web (Google). This figure shows how prominent certain animals are for a Google Search query for [“climate change”]. Animals are scaled by the number of results (in text and image) returned by Google Search. It is found that on the web, for a text query, results are distributed across all the animals and thus do not particularly favor one specific issue animal. Visualization created in Adobe Illustrator, 15 July 2007. Source: Digital Methods Initiative 2007.

Figure 23).[57] Source distance is a technique suitable for PageRank as it makes optimal use of the ranked list presented; the findings may be interpreted as the presence of skeptics in the most authoritative sources on climate change, as ‘status-authored’ by Google which followed the hyperlink network of the web.

In addition to questions about how far from the top a source or keyword is mentioned, the Google result page lends itself to questions of ‘resonance’. Apart from the specific position, ‘resonance’ inquires into how often/in how many sources a specific keyword is mentioned, which can be viewed as an equivalent of ‘air time’ in search engine results. The example project for resonance is ‘issue animals’ which investi-

[57] Research project of the DMI Summer School 2007. Research by Bram Nijhof, Richard Rogers, and Laura van der Vlies.



Issue Animals Hierarchy in the News (Google News)

Research Question_ For the issue of climate change, how prominent is each animal (in text and image)? Are there significant differences per 'sphere' (web, news, blogosphere) in the frequency with which each animal is referenced?
Findings_ In the news, for a text query, the polar bear is the animal most associated with climate change, followed by the cow.

Source_news.google.com
Query_"climate change" + "*animal x*"
Tools_Google News Scraper and Tag Cloud Generator
Date_15 July 2007

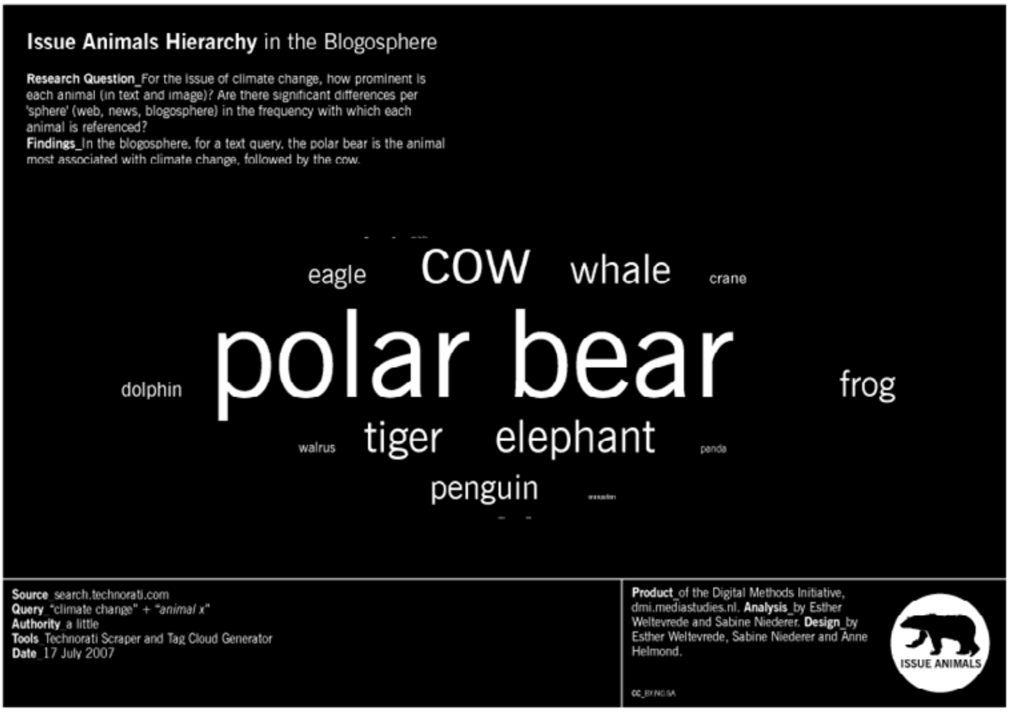
Product_of the Digital Methods Initiative, dmi.mediastudies.nl. **Analysis**_by Esther Weltevrede and Sabine Niederer. **Design**_by Esther Weltevrede, Sabine Niederer and Anne Helmond.

CC_BY.NC.SA

Figure 25: Issue Animals Hierarchy in the News (Google News). This figure shows how prominent certain animals are for a Google News query for [“climate change”]. Animals are scaled by the number of results (in text and image) returned by Technorati. It is found that in the news, for a text query, the polar bear is the animal most associated with climate change, followed by the cow. Visualization created in Adobe Illustrator, 15 July 2007. Source: Digital Methods Initiative 2007.

gated which (endangered) animals resonate most in the climate change result page (see Figure 27). This project did not only look into which animals are mentioned most in the Google Image results, but introduced ‘cross-spherical analysis’ which means that the results of the same technique were compared across different types of Google engines, the Google Web Search and News Search, and the then dominant blog search engine, Technorati (see respectively Figure 24, Figure 25, and Figure 26).[58] In addition to the cross-spherical resonance analysis in text, the project included a cross-spherical resonance image analysis (see Figure 27, only Google Image search results are shown here).

[58] Research project of the DMI Summer School 2007. Research in collaboration with Sabine Niederer.



Issue Animals Hierarchy in the Blogosphere

Research Question_For the issue of climate change, how prominent is each animal (in text and image)? Are there significant differences per 'sphere' (web, news, blogosphere) in the frequency with which each animal is referenced?
Findings_In the blogosphere, for a text query, the polar bear is the animal most associated with climate change, followed by the cow.

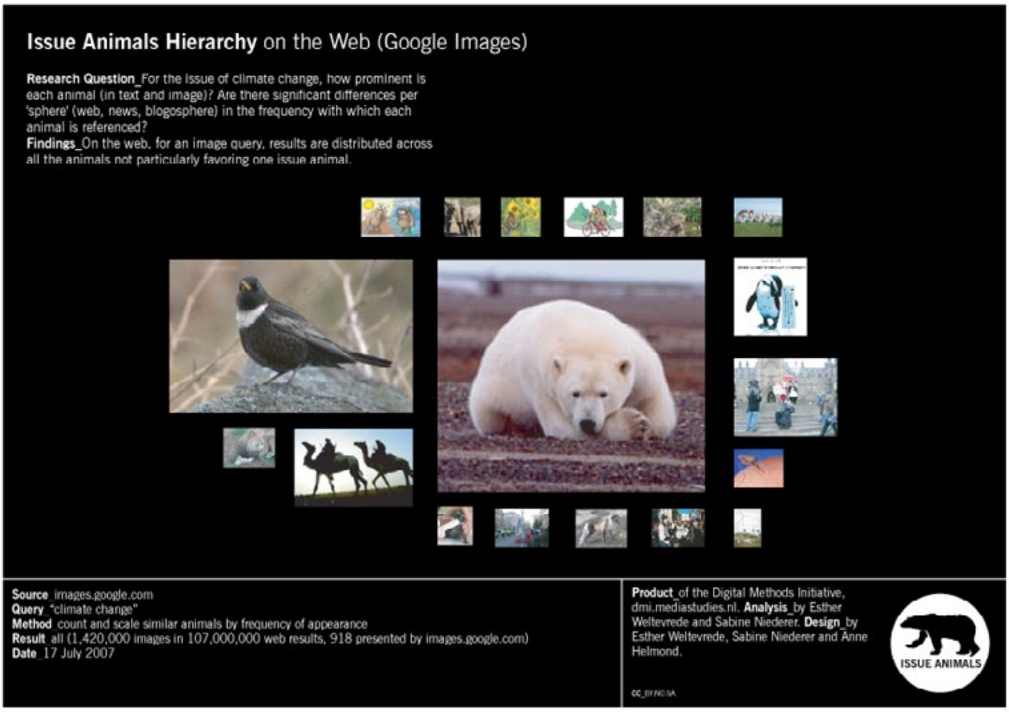
Source_search.technorati.com
Query_"climate change" + "*animal x*"
Authority_a little
Tools_Technorati Scraper and Tag Cloud Generator
Date_17 July 2007

Product_of the Digital Methods Initiative, dmi.mediastudies.nl. **Analysis**_by Esther Weltevrede and Sabine Niederer. **Design**_by Esther Weltevrede, Sabine Niederer and Anne Helmond.

CC_BY.NC.SA

Figure 26: Issue Animals Hierarchy in the Blogosphere (Technorati). This figure shows how prominent certain animals are for a Technorati query for ["climate change"]. Animals are scaled by the number of results (in text and image) returned by Technorati. It is found that in the blogosphere, for a text query, the polar bear is the animal most associated with climate change, followed by the cow. Visualization created in Adobe Illustrator, 17 July 2007. Source: Digital Methods Initiative 2007.

Both projects informed the further development of the Google Scraper, a tool developed by DMI, which batch queries Google and can be used for both source distance analysis and resonance analysis. There are additional options built-in, however, which developed together with the evolution of Google's web search algorithm to which I will return in the following section. In addition to these types of web epistemological research that includes Google in its analysis, as mentioned, the tool is also known as 'the Lippmannian Device'. The Lippmannian Device repurposes Google too, but instead of including for instance the ranking of results into analysis, the research uses of the Lippmannian Device typically draw on editorially selected starting points (URLs) or URLs from another source, and resort to the Google engine and index to query those sources for specific terms. The analysis and findings of this type of research generally do not include findings about Google and can be considered as social research with Google.



Issue Animals Hierarchy on the Web (Google Images)

Research Question_ For the issue of climate change, how prominent is each animal (in text and image)? Are there significant differences per 'sphere' (web, news, blogosphere) in the frequency with which each animal is referenced?
Findings_ On the web, for an image query, results are distributed across all the animals not particularly favoring one issue animal.

Source_images.google.com
Query_"climate change"
Method_count and scale similar animals by frequency of appearance
Result_all (1,420,000 images in 107,000,000 web results, 918 presented by images.google.com)
Date_17 July 2007

Product_of the Digital Methods Initiative, dmi.mediastudies.nl. **Analysis**_by Esther Weltevrede and Sabine Niederer. **Design**_by Esther Weltevrede, Sabine Niederer and Anne Helmond.

CC_BY-NC-SA

Figure 27: Issue Animals Hierarchy on the Web (Google Images). This figure shows how prominent certain animals are for a Google Images query for ["climate change"]. Images are scaled by the number of images returned by Google images. It is found that, except for the polar bear, no animals are particularly 'favored'. Visualization created in Adobe Illustrator, 17 July 2007. Source: Digital Methods Initiative 2007.

A third mode of research that is built on Google as a status-authoring device is the Issue Dramaturg, which can be defined as 'ranking research' by tracking the ranking of results for a specified query over time. The Issue Dramaturg is a tool developed by Govcom.org; it monitors the ranking of sources for a particular query over time, in particular the ways in which Google ranks web sources on its return pages, which are so influential in structuring traffic and 'attention'. The tool displays the constant competition of information sources to be returned in the top ten for any given query. The Issue Dramaturg shows rankings per query, thereby capturing the changing search engine placement of the top results (Govcom.org 2007). The example shown in Figure 28 displays the disappearance of 911truth.org from the Google results for the query [9/11] in September 2007. 'Ranking research' can for instance be used to monitor the volatility of the algorithm by capturing the changes in the ranking of results for a query over time. Analyzing the results for the query [9/11] over time with the Issue Dramaturg, the results of this tool can be used to gauge algorithmic change in an

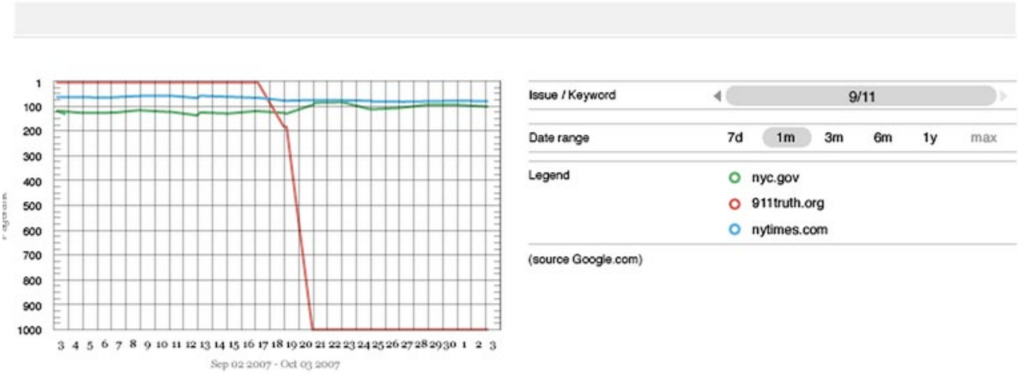


Figure 28: A Website is gone: the apparent removal of 911truth.org from Google results for the query *[9/11]*, September-October 2007. The graph shows how 911truth.org used to be in the top 10 Google results for the query [9/11], but received a dramatic drop starting 16 October, 2007. Source: *Govcom.org 2007*.

auditing algorithm sense. In other words, when analyzing the fluctuations of results over time and correlating changes with known algorithm updates one can extrapolate the consequences of algorithmic changes to a search engine result page for a particular query. The data shows a clear fluctuation in results from 2011 onward. Looking at reported algorithm updates around that time the Panda algorithm change is the most significant, which aims at boosting ‘high quality’ websites and downgrading low quality ones (Borra and König 2013), which as a result privileges the mainstream over the fringe.

‘Source distance’, ‘resonance’, ‘issue research’ and ‘ranking research’ are thus modes of research that have been developed on the basis of the intent behind Google’s PageRank to produce authoritative results. However, almost fifteen years after Page and Brin entered the search market with their PageRank algorithm, the algorithm has changed drastically, turning PageRank into only one of the over 200 signals that Google uses to rank results. The introduction of ever more factors in the algorithm, and the continuous tweaking of how the factors are weighed against each other marked the end of the prospect of obtaining globally universal knowledge through the engine, which is the topic of the following section.

The demise of PageRank: algorithm changes and modes of research

Almost fifteen years after Page and Brin entered the search market with their PageRank algorithm, the algorithm has changed drastically, where PageRank is but one of the over 200 signals that Google uses to rank results, however central (Dean 2015; Sullivan 2010). The implications of what can be known through Google are significant, too, as it is increasingly challenging, if not impossible, to find information with a 'global' relevance, information that is universally recommended. The embedded intent to show the most authoritative has evolved to other assumptions of what are relevant results, such as offering tailored results. The purpose of describing Google's algorithm changes in this section is to address how these changes not only limit what can be known with the algorithm but also how these changes afford new modes of knowledge.

As mentioned earlier, patents are an informative source to understand the assumptions and considerations behind technical changes to the search engine. However, Google may or may not have implemented the patent, or implemented it somewhat differently; thus the patent may not fully correspond to the actual implementation. Patents, however, are interesting because they demonstrate the intentions behind certain changes and suggest which specific signals are used to what end. Connecting patents to algorithm updates—either communicated by Google[59] or reported by SEO and Webmaster communities—helps to understand to what extent patents have been implemented. Hereafter a chronology is provided of the key updates to the search algorithms and their ramifications for digital research using Google as a research device.

Throughout this section the algorithm changes are connected to key updates that in general have received names. Some updates are named by Google (such as Caffeine and Venice); others are coined by various key sources, like SEO communities around search engine updates, including Webmaster World, Search Engine Land, MOZ and Level343. The first name was Boston, given by Webmaster World users at SES Boston. The next few updates—Cassandra, Dominic and Esmerelda—were also named by Webmaster World users, similar in style to how hurricanes are named (Moz.com n.d.). In the beginning, updates often did not receive a name, nor much attention; in fact, the early ones are difficult to fully reconstruct. Since the Boston update in 2003, search engine optimizers, in their quest to get higher rankings for their sites, started

[59] Since November 2011 Google is posting lists of (small and larger) updates monthly or bi-monthly on their "inside search" blog. This is a major change in how Google communicates algorithm changes. For example the first one is described by Matt Cutts (2011).

scrutinizing updates in forums; reverse engineering tried to reconstruct what had changed and assisted with specialized SEO tools such as ‘Rank Trackers’ (Moz.com n.d.). SEO websites are an instructive source for understanding Google algorithm updates as the SEO community is a vigilant, reporting on updates and backing up algorithm changes with screenshots and data.[60]

In 2000 Google hired Amit Singhal, who is currently its senior vice president. One of his first major tasks was to rewrite Page’s original algorithm, to facilitate the inclusion of other ranking signals, turning PageRank from the core algorithm into one of its ingredients, however central. Throughout its history Google has devised ways to accommodate the rapid growth of the web and keep one step ahead of competition and manipulation, making it easier to quickly incorporate new ranking signals. Major change to the system, basically rewriting it entirely, were confirmed twice more with Caffeine and Hummingbird. These major updates facilitated including other signals to the algorithm, initially on a monthly basis, known as the ‘Google Dance’, which was also significant for keeping the indexes in sync across the thousands of servers in Google’s server farm (Sobek 2002). From 2003 onwards it turned into an incremental approach, since an update named Fritz. Google had been critiqued for changing its algorithm continuously both by the trade press (Sullivan 2011) and academia (Helmond 2010). Google CEO Eric Schmidt told Congress that Google made 516 quality improvements in 2010 and tested over 13,000 updates (E. Schmidt 2011); in 2011 this had increased to 520 updates with 58.000 quality experiments (Google 2012). Although the variables in the Google algorithms are over 200, and new tweaks, updates and experiments are constantly tested, the percentage of affected queries range from 0.01—for example, spell check improvements for queries with more than ten terms (Cutts 2012)—to 35%, for example, the 2011 Freshness Update (Singhal 2011); Google expects Hummingbird to affect 90% of the queries (Slawski 2014). The updates included in Figure 29 are a selection of key algorithm updates since they have the biggest impact on search. These include major infrastructure updates but also smaller changes that significantly affect (re)search with Google. Consider for example how the Issue Dramaturg example in Figure 28 can be used for algorithm auditing by studying the implications of larger and smaller updates on the ranking of sources over time.

In what follows, algorithm changes that have resulted in new, or changing, modes of research that were not possible before the change type are discussed. They include ‘Anti-manipulation’ with ‘ranking research’, ‘Personalization / Social’ with ‘personalization research’, ‘Local’ with ‘national web research’, and ‘Freshness’ with ‘realtime

[60] See for example this blogpost by Danny Sullivan on Google algorithm changes related to the ‘miserable failure’ Googlebomb (2007a).

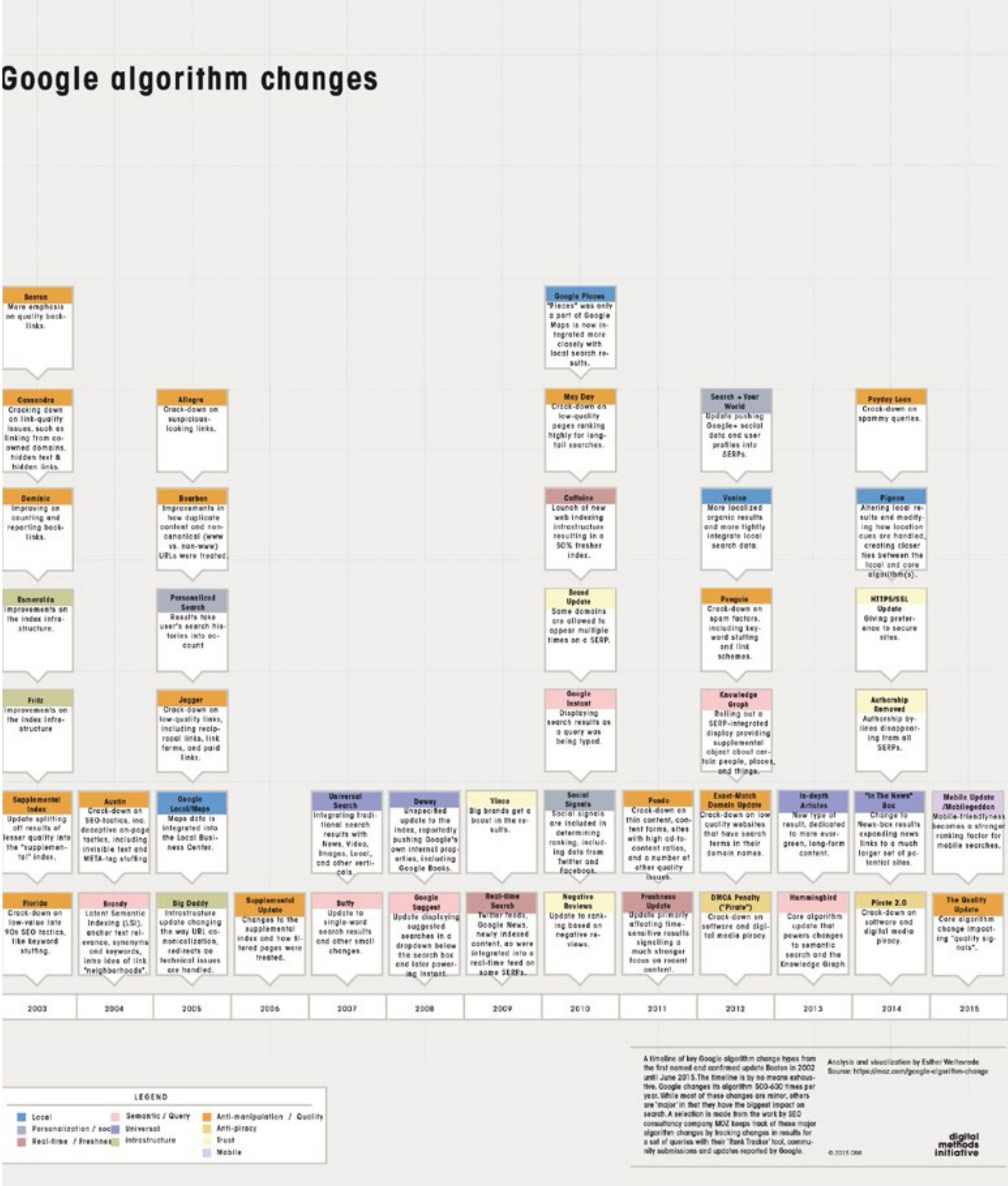


Figure 29: A timeline of key Google algorithm changes from the first named and confirmed Boston update in 2002 until June 2015. The timeline is by no means exhaustive. Google changes its algorithm 500-600 times per year. While most of these changes are minor, others are 'major' in that they have the biggest impact on search. A selection is made from the work by SEO consultancy MOZ, which keeps track of these major algorithm changes by tracking changes in results for a set of queries with their 'Rank Tracker' tool, community submissions and updates reported by Google. Visualization created in Adobe Illustrator, 2015 *(Moz.com n.d.)*. A high-resolution version of this figure is available at *Weltevrede 2015b*.

research'. The algorithm changes, 'Universal', 'Trust (Brands)', 'Anti-piracy', 'Semantic' and 'Mobile' are types that have significantly changed search and research, but have thus far not led to new modes of research under the umbrella of the Digital Methods Initiative and are therefore not further discussed here.

Anti-manipulation and infrastructure

The first and most recurring type of algorithm changes in the history of key Google algorithm updates are designed to combat low-value SEO and spam tactics such as linking from co-owned sites, keyword stuffing, and combating so called 'spammy neighbourhoods'. The dominance of this change type throughout the Google algorithm history demonstrates how influential SEO practices have been in shaping web search as we know it today, a point succinctly made by Andrea Fiorentini (Fiorentini 2014). The anti-manipulation or 'quality' updates have significantly altered search. The type does, however, not permit a societal research type per se but perhaps rather a medium research type, which has a more diagnostic purpose and uses SEO tools and techniques.

Andrea Fiorentini in his thesis on "SEO Matters" takes up the 9/11 truth ranking research in Figure 28 and starts by accounting for the counter-intuitive rankings of 911truth and 911commission in the engine results for the query [9/11], repurposing SEO tools to do so.[61] He finds that the 911truth's citation score is actually higher than the 911commission's score: the trust score, however, is lower. Citation should be considered an impact measure rather than an endorsement measure; the trust score could provide a more reassuring indicator of quality. Moreover, the work looks into the use of the nofollow attribute, which is a link attribute that can be assigned to hyperlinks instructing bots and crawlers not to follow them. The study approaches the nofollow attribute as an indicator of citation value degradation, especially when used by .edu and .gov sites. The nofollow attribute allows .edu's and .gov's in some sense to place a caveat, or comment, on its link to 911truth (Fiorentini 2014).

The term 'reactivity' is useful in this context as it points to how rankings influence what they seek to measure (Espeland and Sauder 2007). SEO'ers will seek to improve their performance, for instance by trying to rank higher in the Google results. When a measure becomes a target, it ceases to be a good measure (Espeland and Sauder 2007).[62] A large number of updates and changes to the algorithm seem intended on

[61] Using one of the most popular backlink explorers in the SEO community by majesticseo.com.

[62] Some metrics seek reactivity as desired effect, such as Klout (Gerlitz and Lury 2014).

countering those who seek to understand the algorithm as a target. Arguably, when Google combats 'reactive' results and optimizes the relevance of results, Google results become more societally relevant for the digital researcher. On the other hand, it does exclude modes of research into spammy neighborhoods.

Personalization, social and personalization research

Google's algorithm increasingly determines what is relevant based on the input from the user, both in terms of the query and in anticipation of what the user wants, based on knowledge of that user. The knowledge of the user may be gleaned at the instant of the query, but also based on already collected knowledge of that user based on his/her profile and on knowledge about statistically and demographically similar users (Beer 2009; Weber and Castillo 2010; Borra and Weber 2012), assembling what Stalder and Mayer call the 'second index' (Stalder and Mayer 2009). Engines like Google not only provide information to users, but also gather knowledge on their users to inform their algorithms. And these algorithms are increasingly dynamic because every query, every click, changes the medium incrementally. For example, if users consistently click on the fourth query result, Google gains in reranking that result to a higher position. Gillespie calls this the 'cycles of anticipation', 'the bits of information that are most legible to the algorithm, and thus tend to stand in for those users' (Gillespie 2014, 173). In this and the following section I will discuss algorithm changes, affecting a very high number of queries and based on knowledge about users: personalization and localization. These two categories of updates supposedly meant the end of a 'global Google' and have repercussions for what can be known with the search engine. A 'global engine' becomes increasingly difficult but is still within reach by configuring the settings, as I will return to later. Localization will be discussed in detail in the following section, here the core personalization updates are discussed.

Google Labs came with Personalized Search in November 2005 (Sherman 2005). At that stage Personalization was primarily based on a bookmark manager, with which a user could bookmark a result by hitting the star next to it (this feature is no longer available). In the corresponding patent 'Methods and systems for personalized network searching' (Badros and Lawrence 2009) it is argued that search engines are mainly used for navigational searches; bookmarking is a way to partly incorporate that behavior in the search engine. Personalization implies that a user's history and clicks on past searches' results get more weight when the users revisits that topic of interest. According to one of Google's product managers, if you search for fish a future search for [bass] will probably be more heavily weighted toward fish than the musical instrument (Sherman 2005). In 2009 Google started offering personalization to everyone, including to users that are not signed-in (Horling and Kulick 2009). Search

History and later Web History became other key signals in personalizing search results (Sullivan 2007b). The search history contains data on queries submitted, results clicked, ads clicked, browsing history, other activities such as email and editing documents, and derived data, such as the time since user visited that page last. In addition, personalization clusters similar users and recommends them similar results (Datar and Garg 2010).

Martin Feuz, Matthew Fuller and Felix Stalder (2011) were among the first to perform empirical personalization research. They trained various 'Google profiles' of French philosophers; they fed the indexes of their books into the engine so as to create Google profiles so that they received results as if they were Foucault. The training sessions were followed by test runs comparing the result pages for the same query across the philosophers' profiles. The findings suggest that personalization does take place to a surprising, albeit trivial extent. A more recent study observed that 11.7% of the results differ due to personalization which is caused by three main factors: 1) being logged in to Google, 2) geography[63], and 3) the search history of the last 10 minutes of searches (Hannak et al. 2013). This type of 'personalization research' can also be considered as empirical 'search research', as it is mainly interested in generating findings about the search engine.

When using the medium to do research, the (advanced) settings of the search engine as well as the browser become important, which may be set so as to disentangle oneself as much as possible from the search engine and configure the search engine as a research device in alignment with the research purpose. For Google this would entail turning off personalization and logging of the search history by configuring the Google settings and creating a 'research browser'.[64]

[63] Here the location of the user is considered as a personalization measure whereas I consider it as a different type of algorithm update since it has allowed for a specific type of research that is addressed in the following section: 'local and national web research'.

[64] The research browser is created as follows:

Step 1. Install a new browser or create a dedicated research profile (for example Firefox)

Step 2. Go to Firefox > Preferences > Privacy

Tell sites 'I do not want to be tracked'

Never remember my history

When using the location bar suggest [nothing]

Save settings

The configuration of the Google settings varies per research project and may include logging out to turn off personalization and location and language settings to retrieve local results.

For a video explaining the steps, see Weltevrede (2015a).

Local and national web research

In 2000 Google started to offer local domain Googles for a growing number of countries or regions, where searches are performed at Google.nl (for the Netherlands), Google.fr (for France), Google.de (for Germany) and so on. By default the engine user is directed to the local domain Google on the basis of their location (through a technique called geoIP which infers the physical location from the IP address of the computer). Whereas the initial dreams of a global cyberspace made Google return global results, we now see what Richard Rogers calls the 'revenge of geography' (Rogers 2012, 1). A query receives local results when the (local) language is inferred from the user's browser (or Google) settings. Since 2000 Google has prescribed what can be known with/through the search engine on the basis of location cues; the way this happens, however, has changed significantly over the years. In this section I give an overview of the most significant changes related to localized search and how this affects what can be known when repurposing Google. Localization has facilitated 'national web research'. The possible scope and specific modes of research have, however, changed significantly over the years.

First of all, localization is not the same as personalization. Personalization was introduced in 2005 and implies that a certain percentage of the results is boosted based on previous searches. Customizing results on the basis of location includes location as one of the ranking factors of the results presented. This chapter includes an overview of significant changes in so-called 'organic search', which appear because of their relevance to the search term but do not include advertisements (see for example Killroy 2005). The changes discussed include differences in results coming from search engines like Google.com, Google.nl, and Google.de, and exclude specialized engines such as Google Local and Google Maps. Over the years, the way in which location is precisely included as a ranking factor has changed significantly. Location was introduced in 2000 with the launch of ten local domain Googles that coincided with the possibility for users to search in their local language, which included French, German, Italian, Swedish, Finnish, Spanish, Portuguese, Dutch, Norwegian and Danish (Google 2000). Nowadays Google is available in 172 local domains in 45 languages (Häsä 2012).

Blogs and patents from the early years of localized search unfortunately do not tell much about the specific ranking factors used in the local domain Googles. In 2012 Google revamped organic localized search with the Venice update, which led to the year being declared the 'year of local' by SEO blog moz.com (Ramsey 2012). Venice, a code-name used by Google, more aggressively localizes organic results. Local search may also be associated with some of Google's patents, most notably about methods to distill 'location intent' in queries with the "Identification of implicitly local queries"

(Diligenti et al. 2012) and “Geographic coding for location search queries” (Buron et al. 2010); and determining the location of web pages, with “Determining geographical relevance of web documents” (Heymans et al. 2011). In the following the assumptions and intent behind localized search are discussed in relation to queries and web pages (see Table 2 for a synopsis).

Website locale	**Local intent query**
	Explicit location in query terms
Content (e.g. contact info)	Implicit local reference in query terms
Local domain	Local domain search engine
IP address	Language
Meta data	IP address
User traffic	Aggregated user behavior
	Use of term in locale

Table 2: Example of signals that are used to identify and define the local intent of queries and the locale of websites. Signals derived from: Diligenti et al. 2012; Buron et al. 2010; Heymans et al. 2011.

Queries can be associated with a user locale. The term ‘locale’ is a poetic word for the place where something happens or is set, and is used by Google to refer to a geographic subdivision of the world, which can be a country or any subdivision of a country (for example state, province, city, district), groups of countries (for example, political unions, groups of countries with a common cultural heritage, countries within particular regions) (Diligenti et al. 2012). This locale can be made very explicit by specifying it in the query, like a search for [restaurant Amsterdam]. Issues that search engines may have to deal with are spelling errors in place names, alternative names for locations and alternative address formats, which are addressed in the patent by Buron et al. (2010). In a more recent patent (Diligenti et al. 2012) not only explicit local queries are assumed to return local results, but also a so-called ‘implicitly local query’ not specifying a location, but where local results might be appropriate, such as a search for [restaurant]. Signals mentioned in the patent (Diligenti et al. 2012) determining whether a query contains an implicit location include first of all the location of the search engine, or put differently, whether the searcher uses the ‘global’

Google.com or one of the local domain Googles. Second, the search engine might look for the implicit local relevance of all query terms, on the basis of their specific local relevance in the past. Third, language may serve as a location signal since some terms have significance in one language but not in another, for example the term ‘tax’ which has meaning in English but not in Spanish; hence it does not carry an ‘implicit preference for local results’ in Spanish speaking countries. Fourth, in addition to language the country associated to the IP address from which a query is issued is significant: terms may have specific local associations, for example the query [freedom] which is fairly generic in most English speaking countries except for Australia where it is also a prominent home furnishing business. Fifth, aggregated user behavior may serve as a signal. When a statistically significant number of searchers using a particular query tend to click on local search results, it is likely they have an implicit preference for local results. In addition, if a statistically significant number of other users have combined a query term with a location in the past, or if an explicit location is added to a follow-up query in the same query session, or if a particular local domain Google is used often, that might also be interpreted as a query with an implicit preference for local results. The same goes for a term frequently appearing in content related to a particular locale.

With the novel signals described in this patent, more and more queries can be classified as containing ‘local intent’, which coincides with some of the implementations in the Venice update, which more tightly integrate local search data. Before Venice, a search for a local service without city modifier would return Maps based local results blended in with national organic results (Buquet 2012). The Venice update triggers Local Universal results in the organic listing, as it is now ‘able to detect when both queries and documents are local to the user’ (Singhal 2012). In the monthly list of updates provided on the Google Inside Search blog, a number of specific improvements related to the Venice update are listed referring to the above mentioned patent. They include the improvement of Universal Search results by better understanding when a query has strong local intent. The more efficient identification of results that are local to the user, ranking them appropriately by understanding which documents are relevant for particular regions or languages and by privileging local URLs over general homepages as a response to queries with a local intent (for example blogspot.ch instead of blogspot.com for users in Switzerland), and improvements to Autocomplete to include suggestions relevant to the user’s country (Nayak 2012).

In addition to patents on queries with a local intent, Google also has patents describing how to determine the locale of web documents (Diligenti et al. 2012; Heymans et al. 2011). Websites may be associated with specific locales through a number of means. First, a location can be contained in the site itself, such as addressing a locale, an address and other contact information. Second, a page may contain business listings that include location information. Third, a page might contain meta data that identifies a location for the site. Fourth, a site can include a country code top-level domain, or an IP address that might link the site to a specific locale. Fifth, the locale of a site may be inferred from the user traffic to a site, such as the click records. However, some sites and pages might be associated with more than one location, all locations, or none in particular. The latter two are considered to be 'globally relevant' (Diligenti et al. 2012; Heymans et al. 2011). Google has confirmed it uses location as an important signal to surface content relevant to a particular country, using algorithms to detect when a website, subdomain, or directory is relevant to a country. In the Venice update the granularity was changed to page level for sites hosting user generated content, local pages of organizations and local navigation homepages (Cutts 2012).

Algorithm changes in some occasions coincide with interface updates that allow the researcher to configure the medium differently. Besides offering local results by default in the local domain versions of the search engine, the results can also be localized via settings. The settings to localize results have changed profoundly as the Region search option in the advance settings was relatively hidden as an option under 'Date, usage rights, numeric range, and more' in historical versions of google.com/advanced_search (Google 2008). More recently, the settings to localize results have moved to the front-page and allow one to access Google as if one is searching from another country. With the exception of Google.com, which is the main engine for the USA, the local search engines can even be configured in such a way to return results as if one was in a specific city. Empirical tests with VPNs confirm that the configured results are indeed similar to the results one receives if one is searching from another location. The exception is the advertisements, which continued to include Dutch results.

Google's localizations make it possible to study the local specificity of issues. In the project The Nationalities of Issues: Rights Types, the word 'rights' was queried in various languages in the relevant local Googles in order to obtain hierarchies of rights types per country (similar to the source distance example explained above, but now for specific locales and in specific languages). Are there distinctive rights that rise to the top in Finland, the Netherlands, France, Italy, Switzerland, Germany, Austria, Sweden, Russia, Japan, Canada, the United Kingdom, Australia, Philippines, Ivory Coast and other countries? (Rogers et al. 2009). The results indicate that countries could be said to have distinctive concerns, according to Google results (Figure 30).[65] For example, ‘everyman’s rights’ in Finland, ‘prostitutes’ rights’ in the Netherlands, ‘computer programmers’ rights’ in Japan and the ‘right to oblivion’ in Italy (the right to have personal data deleted) are unique to the respective countries.

Freshness

The current default Google interface with Instant updates the result page itself in realtime when users type a query, with realtime suggestions for that query with Auto-Complete. Whereas this is a move towards realtime in the front-end, realtime in the back-end of the engine also increased, as described in Chapter 3 about the Politics of Realtime. The introduction of google.com/realtime in 2009 was significant in this respect, as it enabled the search of social media results. It is no longer in effect due to discontinued contracts with the main content provider Twitter (Sullivan 2011). A second significant moment is the Caffeine update to their index, also introduced in 2009. Before the Caffeine update Googlebot agents were sent out to index changes and new content in scheduled intervals, refreshing Google’s main index every week, whereas with the new update the index is refreshed almost instantaneously. In addition to the Caffeine update, Google started to increasingly privilege fresh results over relevant results allegedly to comply with the post-September 11 demand of millions of users for realtime news and trusted websites (Wiggins 2001). Since 2001, Google has changed its algorithm to be a ‘realtime’ search engine for ‘hot’ or ‘happening’ issues privileging fresh results over relevant ones in terms of their PageRank (Wiggins 2001; Singhal 2009; Singhal 2012). The Query Deserves Freshness algorithm is a noteworthy example, as it is developed to determine whether a topic is ‘hot’ and the query may want fresh results in the top of the result page, as in the case of breaking news stories (Singhal in Hansell 2007).

In 2011 another major Freshness Update affected time-sensitive queries. This update was informed by a series of patents of which the ‘Information retrieval based on historical data’ is the first one (Acharya et al. 2008). The patent takes into account the ‘freshness’ of a page as well as the links pointing to it. Without the patent, because of its age a ‘stale’ page may have many links pointing to it, thus obtaining a relatively high PageRank. The Historical Data patent privileges ‘fresher’ pages. The freshness

[65] Project facilitated by the author in the 2009 DMI Summer School project week on The Nationality of Issues: Repurposing Google for Internet Research. Analysis by Vera Bekema, Liliana Bounegru, Andrea Fiore, Anne Helmond, Simon Marschall, Sabine Niederer, Bram Nijhof, Richard Rogers and Elena Tiis, July 13-17, 2009. Visualization created by Vera Bekema and Anne Helmond in Adobe Illustrator, 2009.

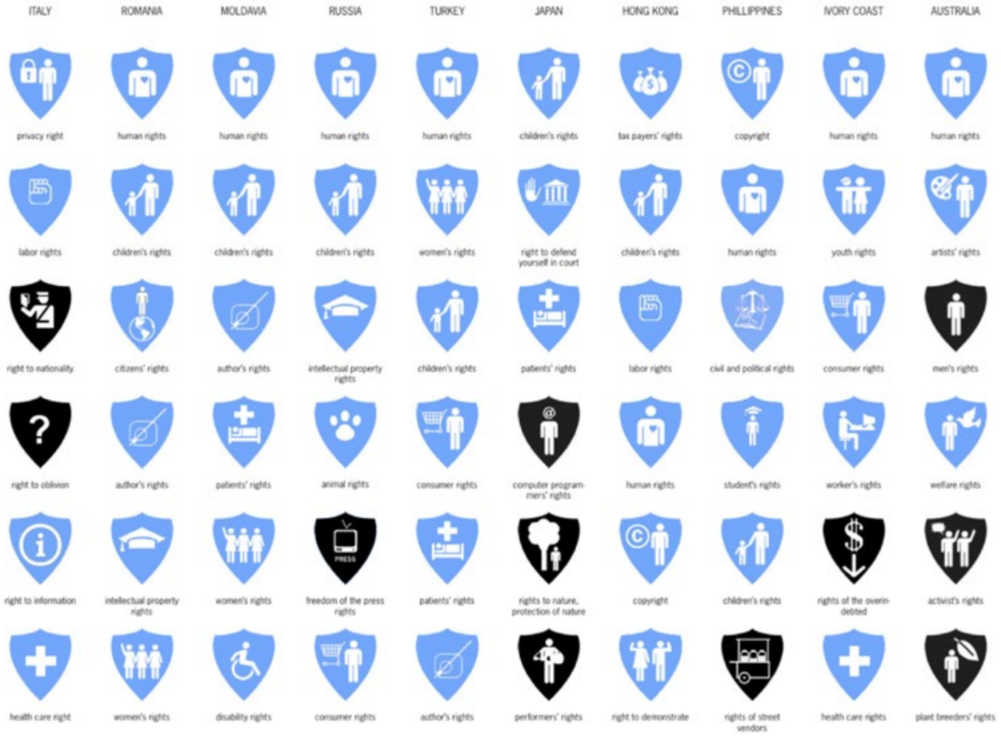


Figure 30: Hierarchies of rights types per country. Top ten distinctive rights types for the query ["rights"] in the local languages of various local Google versions (for example ["oigused"] in Google.ee and ["direitos"] in Google.pt). Results are in the order that Google provided and translated to English. It was found that certain countries have shared concerns, such as human rights, whereas others have unique concerns, such as activist's rights in Australia. Visualization created in Adobe Illustrator, 2009. Source: *Bekema et al. 2010.*

updates have inspired 'realtime research' projects such as 'Historical Controversies Now' (Figure 4), see Chapter 1. The figure displays the top ten results for queries of ten controversies from different historical moments. The results are plotted on a timeline based on their date of publication. The color-coding indicates whether the result's content addresses the controversy historically or whether it is a recent reworking of the event; the shape option indicates whether the issue is still considered controversial. The majority of the results are a year old, more recent results tend to be slightly more controversial, although the vast majority has a neutral tone.

Conclusion

This chapter may be considered as a contribution to the growing field of algorithm studies. I focused on the organizing strategies of the Google algorithm as a basis for what can be known through Google, inquiring into its technicity of organization as a condition for what can be known by using the search engine as a research device. The contribution of this chapter might therefore lay the foundation for the notion of 'search as research' as it paves the way for digital research with the Google algorithms and explores what can be known through them and their development over time. When dealing with a black-box such as the Google algorithms we should not be tempted to disregard them completely for research. Although we cannot inspect its source code, there are myriad methods and techniques to grasp how data is captured and formatted so that we recognize what we may know and how we may repurpose this for digital research. This chapter investigated, grouped, and periodized key changes in the evolution of Google's algorithm, and provided specific 'research settings', in order to find ways to research *with* Google's algorithms.

I have been making a case for methods that are adaptive, sensitive to the changing operational capacities of digital media. In this chapter I argued that specific modes of research are possible not only by focusing on the notion of relevance, but by considering the specific ways in which Google captures, formats, and recommends. The fact that Google is in continuous beta, its algorithms constantly changing, could be seen as an unreliable basis for research with Google. I nuanced this by arguing that updates often only affect a fraction of queries and that larger updates are communicated. This reaffirms the claim that digital research is also always medium research and calls for checking trade press and medium documentation during the research, always keeping a cautious eye when interpreting results for unexpected results. Algorithm updates or Google's discontinuation of specific services such as image search per sphere have a strong impact on repurposing Google as a research device, making certain modes of research impossible. On the other hand, significant algorithm updates have also enabled other modes of research. The evolution of the method of the medium requires an adaptive and flexible approach in order to align with the core concerns and epistemic issues raised by the medium.

Wikipedia’s device culture and the value of dispute

Chapter 6

THE PREVIOUS TWO CHAPTERS engaged with the volatility of both the methods of the medium, as well as with the implications for the digital research methods that follow the research affordances of digital media. As each digital medium is specific and changes continually, the alignment between digital device and research objective is timely and flexible. In this and the following chapter another component in the configuration of the digital device in the research apparatus is considered, namely the specific 'device cultures', or cultures of use and social practices that engage with and are enabled by digital media. The device culture vantage point highlights how digital media are devices that activate and become activated by their uses and practices in digital social life. From a device-driven research perspective, these device cultures can be used to focus the analysis and to find lively, trending, happening, popular, or other focal points relevant for the research objective at hand. In this chapter the device culture of Wikipedia is the object of analysis.

Since its beginning, the free encyclopedia Wikipedia has both been heralded and critiqued for opening up the production of encyclopedic knowledge to the vigilance of the crowds. It was praised as a principal example of 'collaborative knowledge', or 'wisdom of the crowds', where users are empowered as editors and content creators. Others, however, questioned how anonymous editors can provide quality content (Sanger 2009; Keen 2007), and pointed out that articles are susceptible to rapid change, including the arbitrary introduction of misinformation (Chesney 2006; Magnus 2008). Many studies have contributed to the quality debate of Wikipedia's encyclopedic knowledge by comparing articles with other reference works (Giles 2005; Rector 2008), by studying conflict and cooperation (Kittur et al. 2007; Kittur and Kraut 2008; Ekstrand and Riedl 2009), comprehensiveness (Halavais and Lackaff 2008; Clauson et al. 2008; Kittur, Chi, and Suh 2009), the quality of its references (Haigh 2011; Stankus and Spiegel 2010), and bias (Griffith 2007; Hecht and Gergle 2009). Even though concerns about the quality of (some of) the articles persist, Wikipedia's credibility becomes apparent when considering the growing and widespread reliance on Wikipedia by students, journalists, politicians, and so on (boyd 2005; Lanier 2006; Waters 2007).

Although the quality of Wikipedia content has been assessed at length, little attention has been paid to digital research may gain from inspecting the collaborative process of Wikipedia. Following Reagle, who argues that 'reference works can act as "argument engines", sometimes inheriting the conflicts of the external world they seek to document and being seized upon as exemplars and proxies in those debates' (2010, 29), this chapter explores the extent to which Wikipedia's articles, edit histories and talk pages indeed have their cultures of use and social practices inscribed and can be repurposed as valuable historical resources for digital research, beyond the study of the platform itself. By focusing on the device culture of Wikipedia I argue that by studying the process of reaching consensus over controversial topics the substance and dynamics of societal issues may be understood. I introduce the notion of 'device cultures' in digital research to underline how the cultures of use and social practices are key components shaping the purposes of digital devices and how that may be productively repurposed for digital research.

With the notion of 'device culture' I also enter the repurposing debate by empirically and conceptually inquiring into the extent to which the social is inscribed in the web (more broadly) and how it can be made productive in digital research to study societal concerns. In so doing, the chapter again engages with the relation between 'the medium' and 'the social' in device-driven research. I explore this by analyzing the role of the 'device cultures' in digital methods. The device cultures referred to here concern the ways in which the analytical affordances built into the Wikipedia platform both privilege specific practices and uses, and function as key components in negotiations instrumental in producing knowledge. This device cultures perspective is based on the emphasis software studies' take on how technological imperatives of software create the conditions and constraints for its use. Or, as Matthew Fuller phrases it: how software 'forges modalities of experience—sensoriums through which the world is made and known' (2003, 63). And, as touched upon in the previous chapters, by critically looking at the software we can start appreciating what 'grammars of action' are engineered into it (Agre 1994). I connect the notion of 'research affordances' to the notion of 'grammars of action' to conceptualize how the medium prescribes cultures of use and social practices. Phil Agre argues that action is achieved through software, because many of the activities do not exist outside the grammars inscribed in it (1994). Wikipedia and its affordances thus shape the operations and activities enacted through it. The device cultures perspective brings together approaches common in software studies, such as analyzing interfaces and technical documentation, with the 'medium research' approach analyzing digital data to study the device 'in action'. In so doing, device cultures can be rendered productive to study knowledge production on Wikipedia. By drawing attention to how device cultures are specific and immanent to digital devices, I not only focus on the Wikipedia platform's technologies (broadly understood) but take into account the social arrangements and cultural practices they incorporate and enable; I thereby focus on the socio-technical relations producing knowledge on Wikipedia.

Interested in disentangling Wikipedia as a device to be used for digital research, I focus on the various analytical qualities built into Wikipedia and how they may lead to particular research affordances. I will therefore compare the analytical choices with those made in previous Wikipedia conflict and cooperation studies, in order to further disentangle the relation between medium research and social research. The device cultures approach is illustrated by a case study concerning the Wikipedia article on ‘global warming’ by means of Contropedia, a tool to repurpose Wikipedia to study societal controversy. [66] Since its first appearance in 2001, the article on ‘global warming’ has been on Wikipedia’s ‘list of controversial issues’ and since 2006 it is a featured article (Wikipedia contributors 2014f). The article containing the list of controversial Wikipedia articles, also claims that arguments about these types of mature articles tend to reflect the debates of society as a whole and that editing disputes on Wikipedia tend to mirror the controversy’s intensity in the outside world. In this chapter I propose a means by which one can find out what those debates are about and show, at least through the case study, how they reflect societal debates. The main contribution of the device cultures perspective, then, is that it can reveal those disputes’ contents and allows one to map and chart which content within a page was controversial at what point in time.

Wikipedia as a controversy defusing device

As mentioned earlier, device-driven digital research presupposes that engines and platforms have their own specific methods and objects with which digital researchers can work. As stated before, the device cultures perspective takes up the invitation extended by software studies to examine the ‘conditions of possibility’ which software establishes (Fuller 2008, 2) in relation to the social practices and cultures of use it enables and which are inscribed into the device at the same time. It studies how digital culture includes more than human interactions and how it negotiates between software architectures and their users. Platform politics, an emerging area of research, also promotes this reasoning by inviting us to study digital platforms through a combined analysis of both front-end and back-end, taking into account how the platform negotiates different interests (see for instance Langlois et al. 2009; Gillespie 2010; Gehl 2011). Along the same lines, Sabine Niederer and José van Dijck (2010) point out that Wikipedia is a socio-technical device where policy, editors, and content management system (CMS) interplay. In this chapter the focus is on Wikipedia’s device culture, and more specifically, on identifying the medium’s core objectives: how are interests negotiated between Wikipedians and how can we use medium-specific insights to productively repurpose Wikipedia in the context of controversy research?

[66] Contropedia was conceived in collaboration with Erik Borra and is currently further developed in collaboration with a European consortium including Médialab SciencesPo (Tommaso Venturini, Paul Girard, Matthieu Jacomy), Eurecat (Andreas Kaltenbrunner, David Laniado), Density Design (Michele Mauri, Paolo Ciucarelli), and the Digital Methods Initiative (Richard Rogers).

‘Wikipedia’s purpose’, as described in the article dedicated to it, ‘is to benefit readers by acting as an encyclopedia, a comprehensive written compendium that contains information on all branches of knowledge’ (Wikipedia contributors 2014b). In other words, Wikipedia’s aims at ‘encyclopedianess’, the state or quality of being an encyclopedia. This encyclopedianess is defined in terms of exhaustiveness, including all branches of knowledge, as stressed by a quote by one of Wikipedia’s founders, Jimbo Wales: ‘Imagine a world in which every single person on the planet is given free access to the sum of all human knowledge. That’s what we’re doing’ (in Roblimo 2004). Additionally, encyclopedianess is defined in terms of reliability as the other founder, Larry Sanger, states: ‘Our goal with Wikipedia is to create a free encyclopedia; indeed, the largest encyclopedia in history, both in terms of breadth and in terms of depth. We also want Wikipedia to be a reliable resource’ (Sanger 2001). In order to achieve encyclopedianess and whilst upholding one of the first, core, principles that Wikipedia should be open to all, consensus is the primary way to make editorial decisions, as it is the principle guideline for knowledge production and conflict resolution (Wikipedia contributors 2014g). Phrased differently, the different interests of Wikipedians and the potential disputes emerging from them in the content of the articles are negotiated with policy to reach consensus as the main method. To maintain its encyclopedianess, Wikipedia is therefore designed to defuse controversy. In the next section the three key parts of Wikipedia’s device culture are discussed and how these are used to reach consensus.

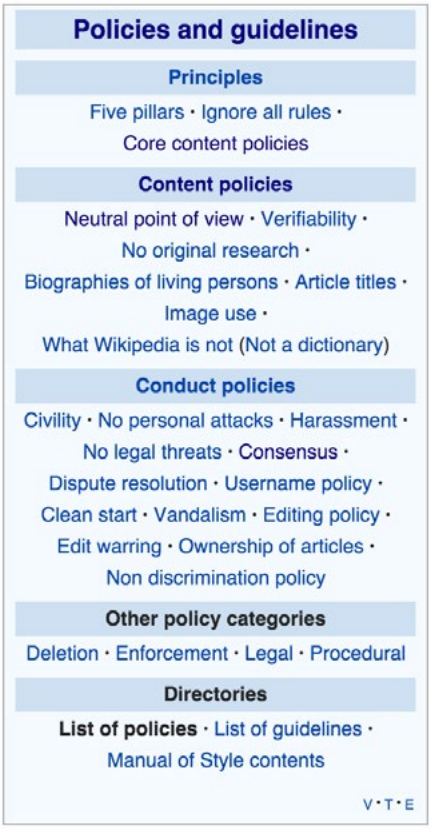
The first aspect of the Wikipedia device culture is architectural, namely its content management system, the MediaWiki software (MediaWiki Contributors n.d.). Like any wiki, it is a web application where anyone can edit content in collaboration with others. A key part of wikis is the use of ‘free links’ markup, where links are written between brackets to refer to other articles relevant in the context of the current one. The Mediawiki software also organizes the front-end (i.e. the encyclopedic content) and the back-end (i.e. edit history and discussion) in a distinct way, by using namespaces, which are prefixes for a given article in order to distinguish between different functions (for example ‘User:’ or ‘Talk:’). Each page thus has an associated discussion page (i.e. ‘Talk:’) where users can discuss the content of the article apart from the article. The edit history page provides access to every version of the article and allows one to compare between two article versions. Although these back-ends are visually complex and difficult to grasp, they are publically accessible and their workings are well documented. The platform is thus relatively transparent and the available data is comprehensive, offering ample research opportunities.

Its policies constitute the second part of Wikipedia’s device culture: principal guidelines for knowledge production and conflict resolution. To maintain its encyclopedianess Wikipedia thus has mechanisms in place to reach consensus, most notably through the three core content policies 'neutral point of view' (NPOV), 'verifiability' and 'no original research' (Figure 31) (Wikipedia contributors 2012b). The NPOV requires that 'all notable and verifiable points of view' are included in an article. The term 'neutral' is somewhat misleading as the policy does not forbid perspectives or views, but rather dictates that articles must not *take* sides, but rather should *explain* the sides, fairly and without bias: Wikipedia seeks to 'describe disputes, but not to engage in them' (Wikipedia contributors 2013d). 'Verifiability' entails that claims and facts must be attributed to reliable external sources as 'readers must be able to check that Wikipedia articles are not just made up' (Wikipedia contributors 2014h). This policy is very closely linked to 'no original research' which states that 'Wikipedia does not publish original thought' and 'all material in Wikipedia must be attributable to a reliable, published source' (Wikipedia contributors 2014d). These policies together support the case that Wikipedia has mechanisms in place to make it a valuable source for cultural and social research as the policies prescribe the norms for practices in knowledge production.

The third part of Wikipedia's device culture consists of the editors continually negotiating editorial changes in the system. When editors are in dispute, they call upon the Wikipedia policy to reach consensus (Niesyto 2011). Editors, also referred to as a Wikipedians, edit Wikipedia; they may be anonymous, in which case an IP address is used in the history pages, registered as a user with a specific username, or an automated tool or bot. Everyone is allowed to edit pages (unless blocked), also without logging in. Logging has advantages, like the ability to create new pages. The tools editors have at their disposal for conflict resolution depends on their level of access. These are hierarchical and range from the blocked user with the fewest permissions to the administrator, or sysop, with high permission levels, who may for instance include locking articles for further editing (Wikipedia contributors 2013c). The various types of editors negotiate within the content management system with reference to the policies in order to comply with Wikipedia's encyclopedianess and, in case of dispute, to seek consensus. In the following, I will consider how one may repurpose these Wikipedia mechanisms meant to achieve consensus in order to study controversies.

Controversy research with Wikipedia

By recognizing that consensus is Wikipedia's primary method to maintain encyclopedianess, scrutinizing the disagreements, apparent throughout an article's edit-history and talk pages where the device culture can be found 'in action', Wikipedia can arguably be repurposed to study controversies. The contribution is positioned in relation



Policies and guidelines

Principles

Five pillars · Ignore all rules · Core content policies

Content policies

Neutral point of view · Verifiability · No original research · Biographies of living persons · Article titles · Image use · What Wikipedia is not (Not a dictionary)

Conduct policies

Civility · No personal attacks · Harassment · No legal threats · Consensus · Dispute resolution · Username policy · Clean start · Vandalism · Editing policy · Edit warring · Ownership of articles · Non discrimination policy

Other policy categories

Deletion · Enforcement · Legal · Procedural

Directories

List of policies · List of guidelines · Manual of Style contents

Figure 31: Wikipedia Policies and Guidelines. Partial screenshot. Source: *Wikipedia contributors 2012b*.

to previous research into conflicts and coordination by addressing the different analytical choices made.

‘Controversy according to Wikipedia’s ‘list of controversial issues’ concerns articles or topics that are subject to edit warring and tend to be biased or have been the subject of NPOV disputes (or will probably be in future) (Wikipedia contributors 2014f). Previous research sought to characterize and visualize such conflict and coordination on Wikipedia, or sought to identify articles which are controversial (Brandes and Lerner 2008; Kittur et al. 2007; Suh et al. 2007; Ekstrand and Riedl 2009; Sumi et al. 2011; Yasseri et al. 2012). Most of these studies are primarily focused on the social dynamics between editors; only a few considered the (types of) content of controversial edits (Viégas et al. 2007; Rad and Barbosa 2012), as is the aim here. So-called edit wars are arguably the most extreme example of actors (thoroughly) disagreeing on Wikipedia.

They occur when authors repeatedly revert each other's edit without resolving the disagreement by discussion, and instead end up in so-called circular editing (Wikipedia contributors 2013a). In Wikipedia, reverting an article to a previous version is a key mechanism for repairing detrimental edits to an article and removes the changes introduced in all intermediate edits (Wikipedia contributors 2013b). Most of the studies set out to identify controversial articles focus on reverts as their prime mechanism to detect whether a page is controversial (see for example Kittur et al. 2007; Suh et al. 2007; Yasseri et al. 2012). Although reverts are generally considered to be a good way to detect controversy, here the focus is not solely on reverts and edit wars; the aim is to include more moderate and subtle disputes about content, too.

The understanding of controversy as put forward here is slightly more nuanced. By aligning the research objective with the medium's research affordances the definition generally used in social theory for controversy mapping followed can be summarized as 'situations where actors disagree (or better, agree on their disagreement)' (Venturini 2010; also see Latour 2005). This definition applied to controversy on Wikipedia can thus be said to be more inclusive as it might also entail disputes not necessarily resulting in edit wars, but for instance turning around the use of a specific adjective or reference. In fact, the analysis showed that most reverts just undo vandalism: deliberate attempts to compromise the integrity of an article by for instance deleting sections, inserting irrelevant obscenities or nonsense. From a controversy perspective, vandalism is not considered a source of disagreement, as it does not introduce a point of view, bias, argument, or opinion. The aim is thus to detect and exclude vandalism by discarding those edits marked as vandalism edits in the edit history.[67] In addition, both punctuation and maintenance edits are filtered out, as well as edits merely inserting new content (i.e. as opposed to deleting or altering it), as changes to existing content convey that an editor disagrees with it and seeks to improve the sentence. After such filtering of the edit history only substantive edits to the content remain. The substantive edit history is subsequently analyzed to indicate which issues have been the loci of most edit activity.

In order to determine around which issues controversy revolves, the device-driven definition of controversy leads me to focus on wiki objects. Consider for example Figure 32, which shows the article on 'global warming' with all plain text blanked out to draw attention to the wiki objects within Wikipedia, such as links to other

[67] It is identified whether a comment contains the word 'vandal', whether the username making the revert belongs to one of the known vandalism bots or when an edit is reverted within 60 seconds). Edits marked as WP:AES are also discarded (Wikipedia contributors 2014a). Note that repeated changes against consensus and edit warring are not regarded as vandalism and that those reverts will not be marked as such.

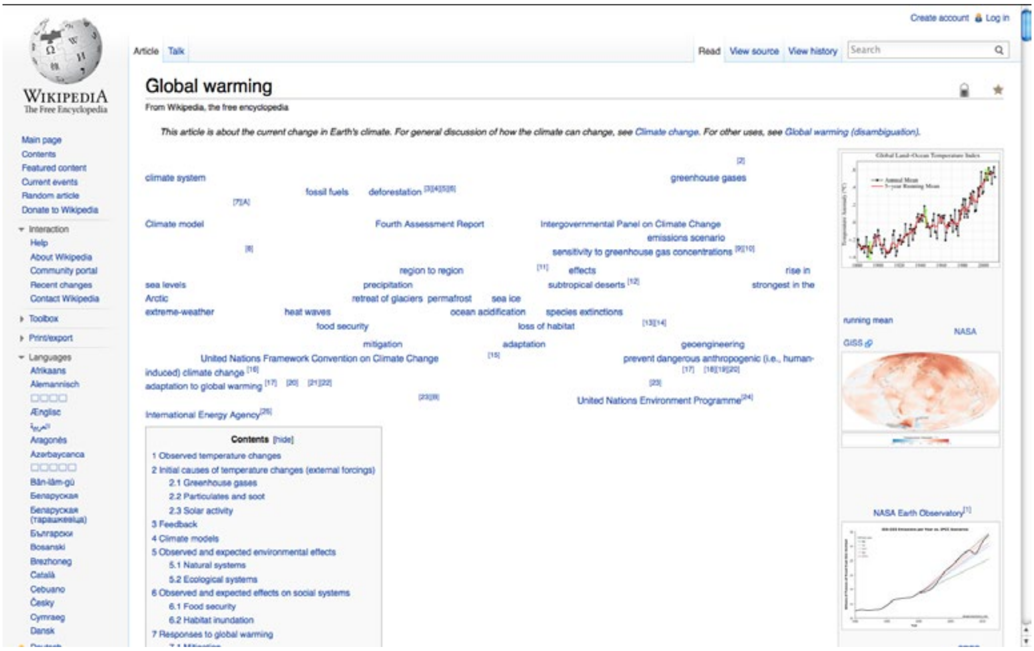


Figure 32: The English Wikipedia article on 'global warming' on January 12, 2013 with all plain text blanked out to emphasize Wikipedia's active content objects. Visualization created in Firefox, 2013. Source: *Density Design et al. 2015*.

articles, references, and images. Wikipedia guidelines indicate that internal links and references should be used to 'increase readers' understanding of the topic at hand'; a phrase should only be linked when it is relevant for the current article (Wikipedia contributors 2014e). Intuitively, for digital social research, wiki objects are important since they all refer to things which themselves are 'matters of concern'.[68] When examining the substantive edits related to an object, the device culture perspective reveals the work put into that object.

[68] Bao et al. (2012) have similarly regarded the links within a specific article as indicative of the subject composition of an article. Through the analysis of common, and unique, links of different language versions of the same article they were able to investigate differing cultural understandings of a topic. Similarly, Hecht and Gergle (Hecht and Gergle 2009) demonstrated on the basis of wiki links that language versions of Wikipedia are self-focused. While in this article we focus on the analytic affordances of internal links, we intend to extend this to other wiki objects such as templates and references.

The controversialness of wiki objects is calculated by counting the substantive edits to the sentences in which these wiki objects appear. However, the more wiki objects appear in an edited sentence, the less the edit focuses on one particular wiki object. For every edited sentence we thus divided the weight attributed to a wiki object by the total number of wiki objects which appeared in that sentence. To find out which wiki object is most controversial, or put differently, around which terms most negotiations took place, the article's wiki objects are ranked in descending order of weighted substantive edit counts. The ranking, then, conveys those wiki objects (issues) yielding most substantive disagreement in the article.[69]

Concluding, the device-driven perspective offers a different perspective to controversy than the definition of controversy used by most Wikipedia studies into controversial articles. Whereas most conflict and cooperation studies focus on the revert as principal indicator for controversy, the alignment between Wikipedia's device culture and the research objective led to the decision to include all substantive topical edits by excluding maintenance edits as well as vandalism, and by clustering the remaining substantive edits per wiki object. The collections of substantive edits per wiki object are clustered in such a way that the relevant edits can be more easily inspected, so that the specific references to policy and discussion threads can be analyzed. In the following I present a case study which makes use of the ideas and measures addressed in the previous section and implemented in an analytical tool.

Case study: Global warming

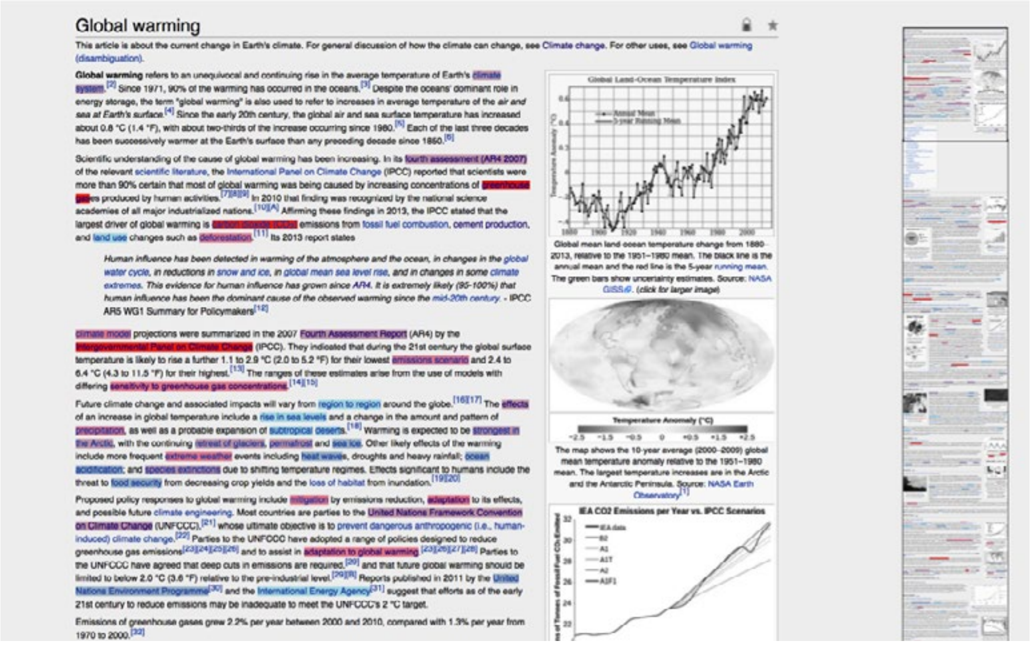
In 2007, the Intergovernmental Panel on Climate Change's (IPCC) Working Group 1 released its fourth assessment report (AR4) (Solomon et al. 2007). It concerned the physical science basis of the natural and human drivers of climate change and concluded that 'warming of the climate system is unequivocal', and that 'most of the observed increase in global average temperatures since the mid-20th century is very likely due to the observed increase in anthropogenic greenhouse gas concentrations' (Solomon et al. 2007, 5–10). Although the report cited thousands of scientists and demonstrated a scientific consensus that the climate is warming and that this is driven by human activity, a minority of scientists opposed the outcomes of the report. Media trying to provide objective information about global warming, faced the danger of being trapped into providing the public with a 'false balance,' by providing equal

[69] For a more elaborate technical description of how Contropedia functions, see (Borra, Weltevrede, et al. 2015).

coverage to both views, thus giving 'disproportionate weight to minority views' (BBC Trust 2011, 66, 58).

Since the publication of the British Broadcast Corporation's (BBC) Trust report in 2011, journalists of the BBC are impelled to pay less attention to climate change skeptics when discussing global warming. The report, which included an independent assessment on the accuracy and impartiality of BBC science coverage by emeritus genetics professor Steve Jones, found that the appropriate application of the editorial guidelines on 'due impartiality' (BBC Trust 2011, 3) was a cause for concern. As this guideline entailed that the impartiality of reporting should be 'adequate and appropriate to the output' and that it should be 'more than a simple "balance" between opposing viewpoints' (BBC Editorial Guidelines), BBC's governing body (BBC Trust) concluded that 'impartiality in science coverage does not simply lie in reflecting a wide range of views, but depends on the varying degree of prominence (due weight) such views should be given' (BBC Trust 2014, 2). Whereas those denying that the climate is warming, or that it is driven by human activity, are very outspoken, they represent a (scientifically) marginal and minority opinion (BBC Trust 2011, 68, 72). Accordingly, BBC journalists should not give them the same attention, or present them as equally valid, as for instance the scientific consensus summarized by the IPCC.

In addition to more traditional mass media Wikipedia is an important source informing the public understanding of science, and public opinion on climate change more specifically. In what follows explore what Wikipedia can add to our understanding of the controversy and more specifically, to focus on how Wikipedians negotiate their differences on the system, in reference to policy. This case study explores whether and how the Contropedia software, through its analysis of the edit history and talk pages of the Wikipedia article on 'global warming', can provide insight into the controversies played out within said article. In particular, the role of the IPCC and its AR4, of scientists denying that global warming is happening at all, and of those refuting that global warming is caused by human activity is analyzed. More generally, we tried to better understand what is, or was, controversial about the article. What are current and past hot topics of dispute? Did dispute about parts of the content for instance focus on specific sub-issues? And if consensus was reached, how was this achieved? With the Contropedia software we tried to condense the edit activity and discussions in the talk pages and explore how this software showcases a controversy's evolution on Wikipedia.



Global warming

This article is about the current change in Earth's climate. For general discussion of how the climate can change, see Climate change. For other uses, see Global warming (disambiguation).

Global warming refers to an unequivocal and continuing rise in the average temperature of Earth's climate system.[2] Since 1971, 90% of the warming has occurred in the oceans.[3] Despite the oceans' dominant role in energy storage, the term "global warming" is also used to refer to increases in average temperature of the *air and sea at Earth's surface*.[4] Since the early 20th century, the global air and sea surface temperature has increased about 0.8 °C (1.4 °F), with about two-thirds of the increase occurring since 1980.[5] Each of the last three decades has been successively warmer at the Earth's surface than any preceding decade since 1850.[6]

Scientific understanding of the cause of global warming has been increasing. In its fourth assessment (AR4 2007) of the relevant scientific literature, the International Panel on Climate Change (IPCC) reported that scientists were more than 90% certain that most of global warming was being caused by increasing concentrations of greenhouse gases produced by human activities.[7][8][9] In 2010 that finding was recognized by the national science academies of all major industrialized nations.[10][A] Affirming these findings in 2013, the IPCC stated that the largest driver of global warming is carbon dioxide (CO2) emissions from fossil fuel combustion, cement production, and land use changes such as deforestation.[11] Its 2013 report states

> *Human influence has been detected in warming of the atmosphere and the ocean, in changes in the global water cycle, in reductions in snow and ice, in global mean sea level rise, and in changes in some climate extremes. This evidence for human influence has grown since AR4. It is extremely likely (95-100%) that human influence has been the dominant cause of the observed warming since the mid-20th century.* - IPCC AR5 WG1 Summary for Policymakers[12]

climate model projections were summarized in the 2007 Fourth Assessment Report (AR4) by the Intergovernmental Panel on Climate Change (IPCC). They indicated that during the 21st century the global surface temperature is likely to rise a further 1.1 to 2.9 °C (2.0 to 5.2 °F) for their lowest emissions scenario and 2.4 to 6.4 °C (4.3 to 11.5 °F) for their highest.[13] The ranges of these estimates arise from the use of models with differing sensitivity to greenhouse gas concentrations.[14][15]

Future climate change and associated impacts will vary from region to region around the globe.[16][17] The effects of an increase in global temperature include a rise in sea levels and a change in the amount and pattern of precipitation, as well as a probable expansion of subtropical deserts.[18] Warming is expected to be strongest in the Arctic, with the continuing retreat of glaciers, permafrost and sea ice. Other likely effects of the warming include more frequent extreme weather events including heat waves, droughts and heavy rainfall; ocean acidification; and species extinctions due to shifting temperature regimes. Effects significant to humans include the threat to food security from decreasing crop yields and the loss of habitat from inundation.[19][20]

Proposed policy responses to global warming include mitigation by emissions reduction, adaptation to its effects, and possible future climate engineering. Most countries are parties to the United Nations Framework Convention on Climate Change (UNFCCC),[21] whose ultimate objective is to prevent dangerous anthropogenic (i.e., human-induced) climate change.[22] Parties to the UNFCCC have adopted a range of policies designed to reduce greenhouse gas emissions[23][24][25][26] and to assist in adaptation to global warming.[23][26][27][28] Parties to the UNFCCC have agreed that deep cuts in emissions are required,[29] and that future global warming should be limited to below 2.0 °C (3.6 °F) relative to the pre-industrial level.[29][B] Reports published in 2011 by the United Nations Environment Programme[30] and the International Energy Agency[31] suggest that efforts as of the early 21st century to reduce emissions may be inadequate to meet the UNFCCC's 2 °C target.

Emissions of greenhouse gases grew 2.2% per year between 2000 and 2010, compared with 1.3% per year from 1970 to 2000.[32]

Global mean land ocean temperature change from 1880–2013, relative to the 1951–1980 mean. The black line is the annual mean and the red line is the 5-year running mean. The green bars show uncertainty estimates. Source: NASA GISS. (*click for larger image*)

The map shows the 10-year average (2000–2009) global mean temperature anomaly relative to the 1951–1980 mean. The largest temperature increases are in the Arctic and the Antarctic Peninsula. Source: NASA Earth Observatory[1]

Figure 33: Controversial issues in the 'global warming' Wikipedia article. Controversialness was calculated on the full edit history until April 16, 2014. The more controversial a wiki link, the redder it is. The images are converted to gray scale. The link to greenhouse gas is among the most controversial. Partial screenshot from the Contropedia Demo tool, 2015. Source: *Density Design et al. 2015*.

Controversialness in the article on 'global warming'

The edit history of the 'global warming' article is analyzed by Contropedia's controversy detection algorithm, as explained in the previous section, and visualized as an additional layer to the article (Figure 33), to provide insight into which issues in the article had received most substantive edits. The visualization is a direct modification of the Wikipedia article, conveying its controversial history; the intensity of the color red of a wiki link indicates the degree of controversy around this issue. The reddest—and thus the most controversial—issues in the lead of the article are 'greenhouse gas', 'carbon dioxide (CO)', 'Intergovernmental Panel on Climate Change', 'climate model', and 'mitigation'. The minified view on the right-hand side of the visualization provide a quick overview of those parts of the article which are most controversial, and show that various sections in the article contain multiple controversial issues.

In addition to the layer view of an article, Contropedia also features a *controversy dashboard* (Figure 34). This output ranks wiki links in an article by their degree of

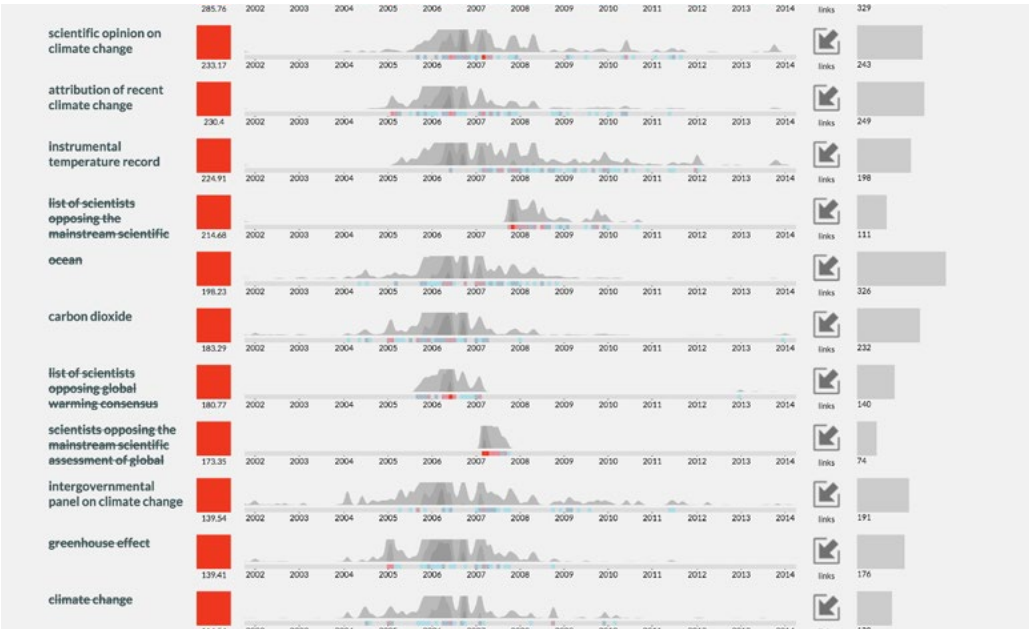

scientific opinion on climate change
attribution of recent climate change
instrumental temperature record
~~list of scientists opposing the mainstream scientific~~
~~ocean~~
carbon dioxide
~~list of scientists opposing global warming consensus~~
~~scientists opposing the mainstream scientific assessment of global~~
intergovernmental panel on climate change
~~greenhouse effect~~
~~climate change~~
285.76
233.17
230.4
224.91
214.68
198.23
183.29
180.77
173.35
139.54
139.41
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
links
329
243
249
198
111
326
232
140
74
191
176

Figure 34: *Partial* dashboard view of the controversial wiki links in the 'global warming' Wikipedia article on April 16, 2014. Each row represents a controversial wiki link. The rows are ordered by how controversial the wiki link is. The redder the square, the more controversial it is overall. Also shown are a timeline of edits and a controversy bar indicating at which time the element was most controversial. Additionally, the type of element and the number of users editing sentences containing that wiki link are shown. The wiki link 'greenhouse gas' is the most controversial. Partial screenshot from the Contropedia Demo tool, 2015. Source: *Density Design et al. 2015*.

controversy. Per wiki link, the dashboard displays additional metrics such as the controversialness of an issue over time, how many times it was edited, and the number of users involved in changes to the object. While the layer view shows the anchor text of wiki links which remained in the article until the moment the controversy detection algorithm was run, the dashboard shows the linked articles possibly removed from the main article. For example, whereas both the article 'list of scientists opposing global warming consensus' as well as 'scientists opposing the mainstream scientific assessment of global warming' were at one point linked from the 'global warming' article, currently only the latter article remains.

The IPCC as a reliable source

One of the top controversial issues is 'intergovernmental panel on climate change' (Figure 35). The top part of the metric 'edits' depicts the edit activity: the higher and

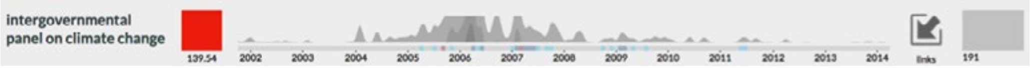
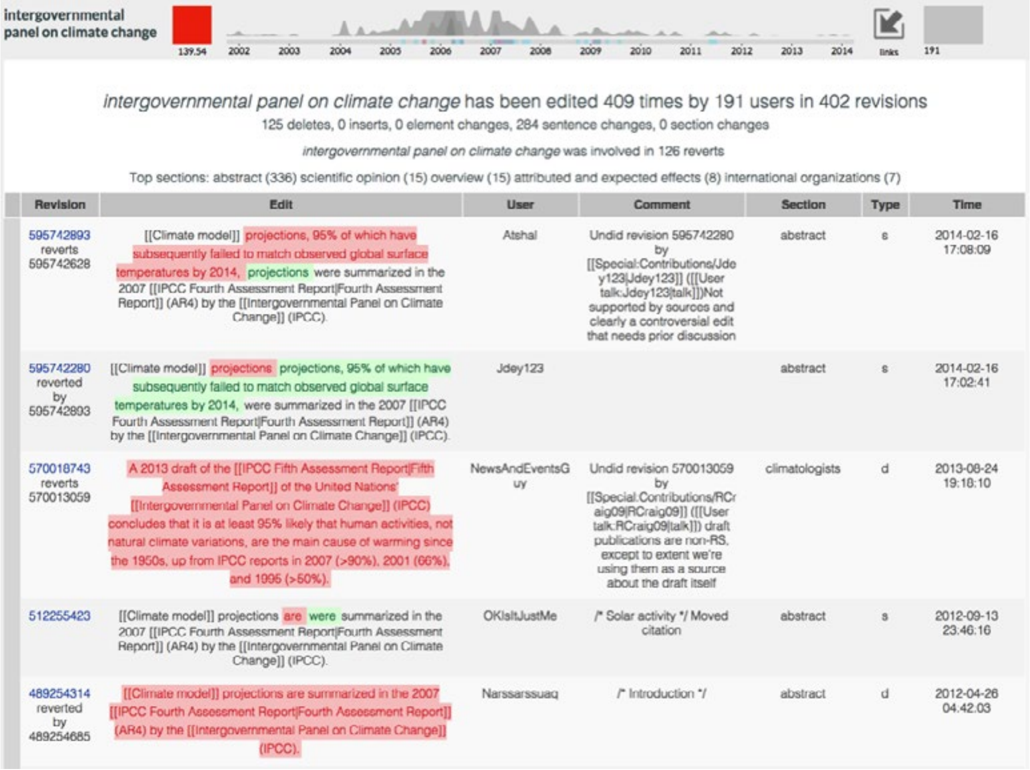


Figure 35: Detail of the dashboard view 'intergovernmental panel on climate change' wiki link in the 'global warming' Wikipedia article on April 16, 2014. The wiki link was most edited within the article on global warming in 2005 en 2007 while it was involved in most edit activity in the middle of 2006. Partial screenshot from the Contropedia Demo tool, 2015. Source: *Density Design et al. 2015*.



intergovernmental panel on climate change has been edited 409 times by 191 users in 402 revisions

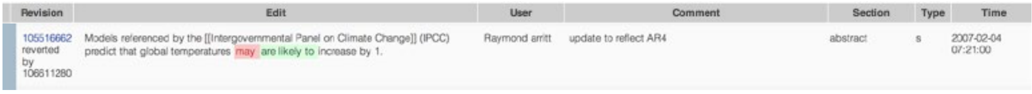
125 deletes, 0 inserts, 0 element changes, 284 sentence changes, 0 section changes

intergovernmental panel on climate change was involved in 126 reverts

Top sections: abstract (336) scientific opinion (15) overview (15) attributed and expected effects (8) international organizations (7)

<table>
<tr><th>Revision</th><th>Edit</th><th>User</th><th>Comment</th><th>Section</th><th>Type</th><th>Time</th></tr>
<tr><td>595742893 reverts 595742628</td><td>[[Climate model]] projections, 95% of which have subsequently failed to match observed global surface temperatures by 2014, projections were summarized in the 2007 [[IPCC Fourth Assessment Report|Fourth Assessment Report]] (AR4) by the [[Intergovernmental Panel on Climate Change]] (IPCC).</td><td>Atshal</td><td>Undid revision 595742280 by [[Special:Contributions/Jdey123|Jdey123]] ([[User talk:Jdey123|talk]])Not supported by sources and clearly a controversial edit that needs prior discussion</td><td>abstract</td><td>s</td><td>2014-02-16 17:08:09</td></tr>
<tr><td>595742280 reverted by 595742893</td><td>[[Climate model]] projections projections, 95% of which have subsequently failed to match observed global surface temperatures by 2014, were summarized in the 2007 [[IPCC Fourth Assessment Report|Fourth Assessment Report]] (AR4) by the [[Intergovernmental Panel on Climate Change]] (IPCC).</td><td>Jdey123</td><td></td><td>abstract</td><td>s</td><td>2014-02-16 17:02:41</td></tr>
<tr><td>570018743 reverts 570013059</td><td>A 2013 draft of the [[IPCC Fifth Assessment Report|Fifth Assessment Report]] of the United Nations' [[Intergovernmental Panel on Climate Change]] (IPCC) concludes that it is at least 95% likely that human activities, not natural climate variations, are the main cause of warming since the 1950s, up from IPCC reports in 2007 (>90%), 2001 (66%), and 1995 (>50%).</td><td>NewsAndEventsGuy</td><td>Undid revision 570013059 by [[Special:Contributions/RCraig09|RCraig09]] ([[User talk:RCraig09|talk]]) draft publications are non-RS, except to extent we're using them as a source about the draft itself</td><td>climatologists</td><td>d</td><td>2013-08-24 19:18:10</td></tr>
<tr><td>512255423</td><td>[[Climate model]] projections are were summarized in the 2007 [[IPCC Fourth Assessment Report|Fourth Assessment Report]] (AR4) by the [[Intergovernmental Panel on Climate Change]] (IPCC).</td><td>OKIsItJustMe</td><td>/* Solar activity */ Moved citation</td><td>abstract</td><td>s</td><td>2012-09-13 23:46:16</td></tr>
<tr><td>489254314 reverted by 489254685</td><td>[[Climate model]] projections are summarized in the 2007 [[IPCC Fourth Assessment Report|Fourth Assessment Report]] (AR4) by the [[Intergovernmental Panel on Climate Change]] (IPCC).</td><td>Narssarssuaq</td><td>/* Introduction */</td><td>abstract</td><td>d</td><td>2012-04-26 04:42:03</td></tr>
</table>

Figure 36: Detail of the dashboard view of the 'global warming' Wikipedia article on April 16, 2014, including a partial edit history. The red color under the 'edit' section indicates a deletion and green an insertion of text. The first few rows of the edit table show how models of global warming are put into doubt both through edit activity and via the comments. Partial screenshot from the Contropedia Demo tool, 2015. Source: *Density Design et al. 2015*.



Revision	Edit	User	Comment	Section	Type	Time
105516662 reverted by 106611280	Models referenced by the [[Intergovernmental Panel on Climate Change]] (IPCC) predict that global temperatures may are likely to increase by 1.	Raymond arritt	update to reflect AR4	abstract	s	2007-02-04 07:21:00

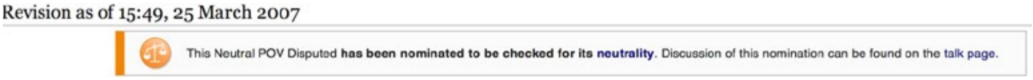
Figure 37: One edit in the history of the wiki link ‘intergovernmental panel on climate change’ in the Wikipedia article on ‘global warming’ April 16, 2014. Red indicates which word was removed and green which words were added. Small semantic changes matter. Partial screenshot from the Contropedia Demo tool, 2015. Source: *Density Design et al. 2015*.

darker the colors, the more edit activity. Most edit activity occurred in 2006 with an additional peak in 2007. In contrast, the small colored bar at the bottom indicates when the element was controversial and to what extent; the redder the bar, the more controversial the issue was at that time. The pace (number of substantive edits per time interval) thus defines the controversy’s intensity. Here, the controversialness bar shows that controversy around the issue sparked in 2005 and that the issue was most controversial throughout 2007. In the following, the edit, and talk page, histories of ‘intergovernmental panel on climate change’ in 2007 are analyzed to find out why the issue was controversial at that time.

Underneath the ‘edits’ tab the substantive edits to the sentences containing the object are listed (Figure 36). When examining the 2007 edits it becomes clear the prominence of IPCC in the article is the main topic of dispute. This is exemplified by changes in sentences like those explaining that global warming is ‘very likely’ due to the observed increase in anthropogenic greenhouse gas concentrations (as mentioned in AR4, see for example Figure 37), and whether the IPCC should be the primary and authoritative source for this claim.

Further examination of some key changes in the edit history in 2007 shows that on 9 January, before the publication of AR4 IPCC, the link to the IPCC is present in the third paragraph of the lead, together with a quote from IPCC’s third assessment report (TAR): ‘Models referenced by the Intergovernmental Panel on Climate Change (IPCC) predict that global temperatures may increase by between 1.4 and 5.8 °C (2.5 to 10.5 °F) between 1990 and 2100’ (Wikipedia contributors 2007a). Following the release of AR4 in February 2007 the sentence was edited and ‘may’ was changed to ‘are likely’, showing the IPCC’s increased conviction (Wikipedia contributors 2007b). However, some editors consider AR4 to be controversial. This can for example be inferred from a revision on 25 March 2007 which attached the point of view (POV) template to the article (Figure 38) and added the comment: ‘This article is nominated for a check’ (Wikipedia contributors 2007c).

Subsequently, in the talk page thread on ‘POV in the intro’ a direct reference was made to the prominence and authoritativeness of IPCC and AR4 (Wikipedia contribu-



Revision as of 15:49, 25 March 2007

This Neutral POV Disputed **has been nominated to be checked for its neutrality**. Discussion of this nomination can be found on the talk page.

Figure 38: POV template on the ‘global warming’ Wikipedia article on March 25, 2007. Partial screenshot. Source: *Wikipedia contributors 2007a*.

tors 2010a, 24). Some Wikipedians thought that the IPCC figured too prominently in the lead; the discussion revolved around the extent to which this was in line with NPOV policy as well as the extent to which opposing views were significant enough to be taken into account following the ‘due and undue weight’ policy. This Wikipedia policy is part of the NPOV policy and requires that each article ‘fairly represents all significant viewpoints that have been published by reliable sources, in proportion to the prominence of each viewpoint in the published, reliable sources’ (Wikipedia contributors 2014c). In a second talk page thread about the POV template in 2007, the viewpoints on global warming were further discussed in relation to how it was phrased in the lead in March 2007 as: ‘this conclusion has been endorsed by numerous scientific societies and academies of science, a few scientists disagree about the primary causes of the observed warming’. By the end of 2007, the POV template was removed and the link to the IPCC had moved up to the first paragraph of the lead. Following the discussions, vague terminology such as ‘numerous’ and ‘few’ had been replaced by concrete numbers and organizations which had been verified with reliable sources indicated by the references in footnotes:

> The Intergovernmental Panel on Climate Change (IPCC) concludes “most of the observed increase in globally averaged temperatures since the mid-20th century is very likely due to the observed increase in anthropogenic greenhouse gas concentrations”[1] via the greenhouse effect. [...] These basic conclusions have been endorsed by at least 30 scientific societies and academies of science, including all of the national academies of science of the major industrialized countries. While individual scientists have voiced disagreement with some of the main conclusions of the IPCC, the overwhelming majority of scientists working on climate change are in agreement with them.[4] (Wikipedia contributors 2007d).

An inspection of the substantive edit history and talk pages bring to light that the publication of IPCC’s fourth assessment report in 2007 was a strong drive not only for the prominence of authoritative sources in the lead but also for changes to sentences containing a link to the article on the IPCC. Wikipedia’s ‘due weight’ policy, and its consideration on how to deal with the different views proportionately, much resembles with the advice to the BBC on how to deal with scientific opinion as ‘due impartiality’ (as opposed to ‘balanced’). However convincing Wikipedia’s ‘due weight’ guideline may seem in theory, for Wikipedia editors it is not immediately evident how

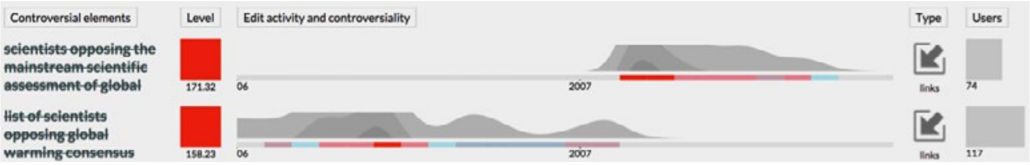

Controversial elements
Level
Edit activity and controversiality
Type
Users
~~scientists opposing the mainstream scientific assessment of global~~
171.32
06
2007
links
74
~~list of scientists opposing global warming consensus~~
158.23
06
2007
links
117

Figure 39: Detail of the dashboard view of the 'global warming' Wikipedia article from February 1, 2006 to February 1, 2008. The second wiki link was used and controversial in 2006, whereas the first wiki link was present in the article throughout most of 2007. The names of the wiki links are struck through as they do not appear in the article on February 1, 2008. Partial screenshot from the Contropedia Demo tool, 2015. Source: *Density Design et al. 2015*.

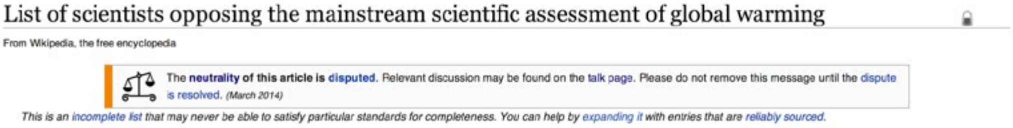
the different views should be weighed in practice. I will elaborate on this in the next section.

Measuring 'undue weight'

The analysis of the related edit history and talk pages suggests that the publication of AR4 ignited controversial edit activity, as well as discussions around the wiki link 'Intergovernmental Panel on Climate Change'. The analysis also indicates that the report's publication affected other parts of the lead. In this section Contropedia's historical view is used to showcase how the controversialness of issues may change over time, focusing on the substance of the edits to the 'global warming' article, before and after the release of AR4.

The most controversial issue in 2006 was 'list of scientists opposing global warming consensus' while in 2007, when AR4 was released, the most controversial issue was 'scientists opposing the mainstream scientific assessment of global warming.' While the content of both articles was in fact the same, its title changed from the former to the latter in early 2007. Figure 39 shows that the controversial edit activity increased after the publication of AR4.

The edit history of both articles suggests that the discussion revolved around the degree of support for AR4 in the academic community. The main dispute, as seen in the substantive edits, turned around the size of the group of scientists opposing global warming consensus (or mainstream assessment); it had been qualified as 'a few', 'relative few', 'other', 'a small number of', 'many', 'a number of', 'about two dozen', 'a small minority of', 'a growing minority' and so on. Wikipedia policy calls such words 'weasel words' which are defined as 'words and phrases aimed at creating an impression that something specific and meaningful has been said, when in fact only a vague or ambiguous claim has been communicated' (Wikipedia contributors 2014e). The



List of scientists opposing the mainstream scientific assessment of global warming

From Wikipedia, the free encyclopedia

The **neutrality of this article is disputed**. Relevant discussion may be found on the talk page. Please do not remove this message until the dispute is resolved. *(March 2014)*

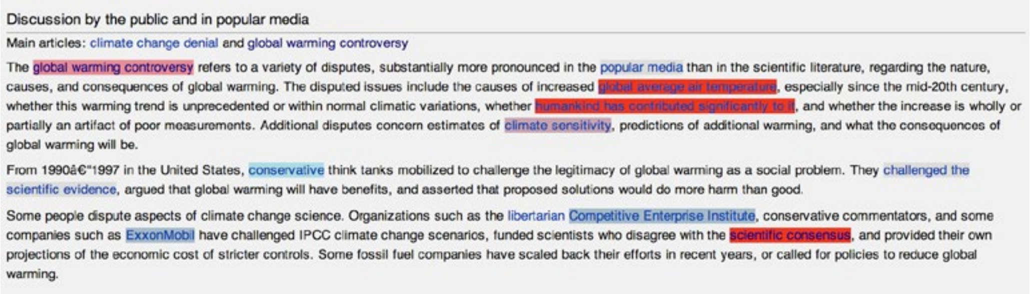
This is an incomplete list that may never be able to satisfy particular standards for completeness. You can help by expanding it with entries that are reliably sourced.

Figure 40: Disputed neutrality template on the Wikipedia article 'List of scientists opposing the mainstream scientific assessment of global warming'. Partial screenshot. Source: *Wikipedia contributors 2007e*.

policy advises that vague terms should be rewritten, supported with reliable sources, or tagged with the appropriate weasel word template. Wikipedia thus provides an interesting place to study how the 'undue weight' guideline is applied in practice, as the 'weasel word' policy, among others, guide Wikipedians to be as concrete as possible by supporting claims with reliable sources.

When looking at the article 'list of scientists opposing global warming consensus', it shows that it seeks to name and list those opposing the consensus of anthropogenic causes of global warming. Wikipedians thus dealt with weasel words in the 'global warming' article by linking to an article detailing the vague qualifier. The article further divided the list into categories such as those that: 'believe global warming is not occurring or has ceased', 'believe accuracy of IPCC climate projections is inadequate', 'believe global warming is primarily caused by natural processes', 'believe cause of global warming is unknown', and 'believe global warming will benefit human society' (Wikipedia contributors 2007e). It is worth noticing that within the most controversial link in 2007—'scientists opposing the mainstream scientific assessment of global warming'—its NPOV was disputed quite often because the list could never be complete (Figure 40), and because the name of the article had been changed several times. Generally, it is discouraged to change article titles unless there is a good reason for it; consensus should be reached about the change. The article had been nominated for deletion a number of times and a name change had been suggested in these discussions. The title was relatively long and could be said to reflect that the controversy around it had not yet resolved (Wikipedia contributors 2010c, 11). The archived discussion threads (including the one 'title discussion') from the beginning of 2007 very prominently featured AR4 (Wikipedia contributors 2010d, 4). As the term 'consensus' was considered controversial in the previous title, in the end the editors reach consensus over a name that referred to the 'mainstream scientific assessment'. As one editor explained, 'mainstream scientific assessment' was preferred because it 'echoes the "assessments" that have been formally done; for example TAR, AR4, U.S. National Assessment, and so on' (Wikipedia contributors 2010e, 4).

Returning to the problem of determining the size and weight of the different views, the tactic of Wikipedians was to link it to a dedicated article seeking to name and list



Discussion by the public and in popular media

Main articles: climate change denial and global warming controversy

The global warming controversy refers to a variety of disputes, substantially more pronounced in the popular media than in the scientific literature, regarding the nature, causes, and consequences of global warming. The disputed issues include the causes of increased global average air temperature, especially since the mid-20th century, whether this warming trend is unprecedented or within normal climatic variations, whether humankind has contributed significantly to it, and whether the increase is wholly or partially an artifact of poor measurements. Additional disputes concern estimates of climate sensitivity, predictions of additional warming, and what the consequences of global warming will be.

From 1990â€"1997 in the United States, conservative think tanks mobilized to challenge the legitimacy of global warming as a social problem. They challenged the scientific evidence, argued that global warming will have benefits, and asserted that proposed solutions would do more harm than good.

Some people dispute aspects of climate change science. Organizations such as the libertarian Competitive Enterprise Institute, conservative commentators, and some companies such as ExxonMobil have challenged IPCC climate change scenarios, funded scientists who disagree with the scientific consensus, and provided their own projections of the economic cost of stricter controls. Some fossil fuel companies have scaled back their efforts in recent years, or called for policies to reduce global warming.

Figure 41: 'Discussion by the public and in popular media'. Section of the 'global warming' Wikipedia article on April 16, 2014. The more controversial a wiki link, the redder it is. Climate change denial and the global warming controversy are described as a media discussion. Partial screenshot from the Contropedia Demo tool, 2015. Source: *Density Design et al. 2015*.

the various scientists and scientific bodies in the various views supported with reliable sources, for example in the 'list of scientists that oppose the mainstream assessment of global warming'. Linking to a dedicated article functions both as a solution to dealing with weasel words, without adding too much detail in the main article, as well as a pressure valve for the 'global warming' article by relocating controversy elsewhere. The analysis of the substantive edit history and talk pages in the year after AR4 suggests that the report itself was a 'reliable source' in Wikipedia policy and helped to make the article more substantive with authoritative sources. Whilst the global warming consensus got a name, it sparked opposition to the authority by providing lists of those opposing the mainstream, also supported by reliable sources.

Global warming consensus?

On 16 April 2014, at the time of the case study, the 'global warming' article did not mention opposing views in the lead or links to the article. Controversy about global warming had been moved to one of the bottom sections of the article titled 'discourse on global warming' and included the sub-sections 'political discussion', 'scientific discussion' and 'discussion by the public and in popular media'. Each sub-section contained links to dedicated pages discussing the specific controversy or debate in more detail. It is most noticeable in reference to the introduction of this case study, that in this Wikipedia article climate change denial and the global warming controversy were described as a media discussion, ignited by conservative think tanks and others (Figure 41).

The link to the 'list of scientists opposing the mainstream assessment of global warming' had however disappeared from the article altogether. The final edits to the link at the end of 2009 again redirected to the talk pages, where significant discussions were taking place about the neutrality of the article and about keeping balance in the scientific views on global warming. In reference to the link one of the editors argues:

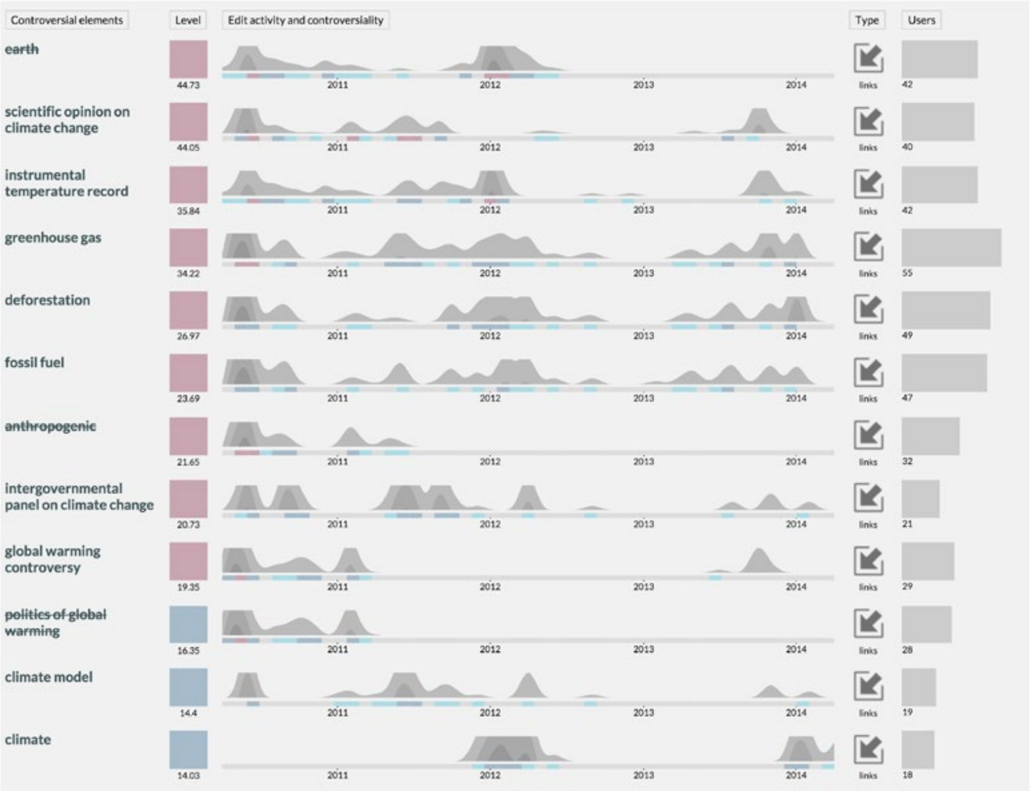
> I am seriously opposed to having that list of 'scientists' in the lede of this article. It has been added to the final section in an appropriate manner by another editor, and that should be the only appearance it makes. Bear in mind that the skeptics represent a tiny majority of scientists - so tiny, in fact, that their views should be considered to be on the fringe. Giving them a 'voice' in the introduction would be a gross violation of WP:WEIGHT (Wikipedia contributors 2010b).

Again, 'undue weight' is the key policy element referenced here. The gist of the consensus reached was that 'in order to be included, the "sides", or more accurately, "views" [...] must be shown to be significant' (Wikipedia contributors 2010b, 55). Around 2010 controversy on the 'global warming' article seemed to have been resolved or at least cooled down, as indicated by the decrease in (controversial) edit activity. This is further supported by focusing the dashboard view of the 'global warming' article to the last four years where the wiki links are hardly controversial (Figure 42). Although the same controversial links still remained present in the top 10, it also shows that new issues had appeared and others had become more controversial. The decline of controversial edit activity in the last four years of the article indicates that the controversy around this link to the list of opposing scientists had cooled down.

Conclusion

Tapping into the repurposing debate and the connection between the study of the medium and the social, I examined how Wikipedia's device culture—prescribed into the Wikimedia CSS and policies, and inscribed into an article's edit history and talk pages—can be used to study the dynamics of societal issues. Treating the online encyclopedia as a device to do digital research with, my analysis of the substance and trajectory of controversies in the article on global warming opens up a series of contributions to both Wikipedia studies and social research with Wikipedia.

Following insights from software studies, and following medium method, I established Wikipedia's 'conditions of possibility' (Fuller, 2008) for using it as a controversy exploration device. Wikipedia's purpose and guidelines to produce encyclopedic content in an open system in which everybody is allowed to edit, combined with a back-end where the process of reaching encyclopedianess is meticulously logged and available for public scrutiny, makes one recognize that the platform is designed



Controversial elements	Level	Edit activity and controversiality	Type	Users
~~earth~~	44.73		links	42
scientific opinion on climate change	44.05		links	40
instrumental temperature record	35.84		links	42
greenhouse gas	34.22		links	55
deforestation	26.97		links	49
fossil fuel	23.69		links	47
~~anthropogenic~~	21.65		links	32
intergovernmental panel on climate change	20.73		links	21
global warming controversy	19.35		links	29
~~politics of global warming~~	16.35		links	28
climate model	14.4		links	19
climate	14.03		links	18

Figure 42: Partial dashboard view of the controversial wiki links in the ‘global warming’ Wikipedia article between April 16, 2010 and April 14, 2014. In this period for example scientific opinion on climate change is more controversial than greenhouse gas. Partial screenshot from the Contropedia Demo tool, 2015. Source: *Density Design et al. 2015*.

to facilitate consensus about content. Recognizing the method to reach consensus as a research affordance to study controversy with Wikipedia, disagreements can be scrutinized in the back-end, like the substance and evolution of the global warming controversy shows.

Additionally, I focus on the analytical choices made in Wikipedia conflict and cooperation studies as well as in controversy research with Wikipedia. Aligning the device culture’s medium-specific characteristics productively in the context of controversy research, edits such as vandalism and maintenance edits are filtered out, which are typically taken into account in Wikipedia conflict studies that seek to research the platform’s resilience to malpractice. By excluding the edits that do not exceed the boundaries of the platform in terms of reflecting societal controversies, substantive edits to wiki objects can be collected per article. The resulting collections allow for two qualitative analytic steps: firstly, to read and analyze varying controversies within an article, and secondly, to discern the entangled roles of policy, CMS and editors in dealing with and resolving controversies. Future work may include developing this qualitative work by for example including templates (such as article locks, point of view problems, reference needed) in the list of substantive edits and by relating discussion threads with wiki objects in the article.

The analysis of the negotiations between Wikipedians shows how the device-driven perspective links the medium-specificity of the platform with the social arrangements and cultural practices the platform incorporates and enables. The case study, for example, underlines the prominent role of the 'due weight' policy in resolving edit disputes. At least some of the policies can be recognized as also having value outside Wikipedia, whereas these policies are designed to include external (reliable) sources, as one of the core content policies forbids 'original research' and instructs Wikipedians to build on 'reliable sources'.

The case study shows that the device-driven approach to focus on wiki links may indeed provide insight into the substance of controversies. Furthermore, it leads to the conclusion that the Wikipedia edit history can direct us towards relevant societal controversies in the debates around global warming. In the context of controversy research the device culture perspective to Wikipedia not only provides a rich account of the practical application of editorial guidelines within the Wikipedia platform, but it also functions as a rich historical reference work for a detailed collection of important scientific and media publications, as well as other events that played a role in the heating up, cooling down, displacement and solution of controversies. The current form of the edit history and talk pages, is such, however, that they are very difficult to analyze. The Contropedia software, which is open-sourced, contributes to the use of Wikipedia as a rich tool for social researchers interested in the historical dynamics of any controversy with a significant number of edits on Wikipedia, as it delivers only substantive edits and recognizes the role of wiki objects.

The national web across devices

Chaper 7

IN THE PREVIOUS CHAPTER the notion of device cultures in digital research was introduced to highlight how cultures of use and social practices are key components in digital devices and how that may be productively repurposed for digital research. Device cultures may be defined as the interaction between user and platform or engine, the data collected, how they are analyzed, and ultimately the resulting recommendations. Subsequently, this chapter offers a comparative device culture approach to studying the national web across a variety of locally popular digital media. I empirically and conceptually inquire into the extent to which the web (more broadly) embeds the social and how it can be made productive in digital research to study societal concerns. In the case study reported on, device cultures are made productive for digital research to sample a data set of relevant and timely URLs for a specific web. This data set of URLs is subsequently studied on the basis of certain common measures (such as responsiveness and page age) to make findings about the liveliness of a web and the extent to which it is censored.[70] The selection technique offered by comparative device culture analysis thus contributes to censorship research.

In (2007) Ricardo Baeza-Yates and colleagues at Yahoo! Research in Barcelona published a review article on “characterizations of national web domain” (Baeza-Yates, Castillo, and Efthimiadis 2007). They sketched an emerging field, which I will refer to as ‘national web studies’. The distinction the authors made in the article between studies in the 1990s on the characteristics of the web and those a decade later on national webs (Kehoe et al. 1999; Baeza-Yates, Castillo, and Efthimiadis 2007) were particularly interesting.[71] This rise of the national was also clearly present in the evolution of Google Web Search as described in Chapter 5. The term ‘national web’ is useful for capturing a historical shift in the study of the Internet, and especially how the web’s location-awareness repositions the Internet as a study object. A national web is one means of summing up the transition of the Internet from ‘cyberspace,’ which evokes a placeless space of email and packets, to a web of identifiable national domains and

[70] Liveliness was introduced in Chapter 2 as an attribute of content dynamics, in this chapter liveliness is measured in terms of response codes measuring whether or not a web page is online.

[71] Further literature on national webs includes the pioneering ethnographic study of the national web of Trinidad and Tobago, where Trini rather than global culture is featured, as well as well-known works on media as organizing national sentiment and community more generally (Higson 1989; B. Anderson 1991; Miller and Slater 2001; Ginsburg, Abu-Lughod, and Larkin 2002). In policy studies, too, national webs, or portions of them, have been ‘mapped’ to inform debates about the extent to which the web, and especially the blogosphere, organizes voice (J. Kelly and Etling 2008; Etling et al. 2010). Related work building tools to circumvent censorship is also interesting, so that voices can still be heard (Glanz and Markoff 2011; Roberts, Zuckerman, and Palfrey 2011).

websites whose content, advertisements and language are matched to one's location. As location-awareness is embedded into a wide variety of current dominant web technologies, these media can be operationalized to study the web nationally. It enables the study of the current conditions of a web space demarcated along national lines, as Baeza-Yates and colleagues pointed out in comparing several national webs. As it is dealt with here, it may also be useful for the study of conditions not only of the digital, but also how it relates to social reality on the ground.

Building upon the web characterization work, this device cultures approach to the study of national webs both engages a series of methodological debates (how to study a national web) and provides an overall rationale for their study (why study a national web). In case of the former, the approach takes into account the multiplicity of user web cultures as well as the related web data collection practices (that users may actively or passively participate in). Search engines and other web data companies such as Alexa routinely collect data from users who, for example search and use their toolbars. Platforms where users share by posting and by rating are also data collection vessels and analysis machines. The outcomes of these data gathering and counting exercises are often ranked URL lists, recommended to other users. When location is added as a variable, the URL lists may be country or region specific. The same holds for language: websites entirely or partly in a particular language are served. This in practice leads to country-specific and/or language-specific webs organized by the data collected and analyzed by engines, platforms and other digital media.

In the case study below, a series of digital media and the kinds of national webs they organize are discussed. Moreover, with the notion of device culture I am specifically interested in the way the users of each of the digital media engage with content, how that engagement is subsequently quantified, measured and weighed to sort, rank and recommend URLs in a specific way. In other words, the notion of device culture allows me to repurpose the socio-technical apparatus of algorithmic media for research. The resulting URLs can be considered to be 'hot' (Likekhor), 'popular' (Balatarin), 'relevant' (Google), 'top sites' (Alexa), or represent a target 'audience' (Ad Planner); all of which are compiled by the interplay between algorithms, affordances and user engagement embedded in the different media. The resulting URLs are each formed by the engines and platforms collecting their data and ultimately intended to represent or provide a country-specific and/or language-specific web, in whatever way. Put differently, the focus is on location-aware media—not only collecting but also serving web content territorially (which is usually nationally) or to a particular language group. As addressed more in depth in Chapter 4, I found that users' and web-content locale may be captured in different ways by digital media, for example by IP-address, language or more implicitly from content and the users' intent connected to a locale. Locale has therefore at least those two distinct meanings, which in certain cases can be reconciled and in other cases not. An engine may serve language-specific websites originating from inside the country as well as from outside the country concerned. For example, the return to a query to the Brazilian 'local domain Google' (Google.com.br) may just as well come from Portugal as from Brazil, both being in Portuguese. So when discussing the dwindling of cyberspace, and the rise of a location-aware web, tension arises between two new dominant ways of interpreting the study object: national webs versus language webs. I am sensitive to that tension, and aware that 'the local,' which as mentioned is how Google refers to its national domain engines, may refer to a national web, a language web or both.

I also discuss how a national web may be demarcated, including the relevant selection procedures. I am particularly interested in the fruitfulness of research outcomes from both separating as well as triangulating the various parts of a national web. Are the URLs listed as 'hot' by blog aggregators similar to the URLs listed as 'popular' by crowd-sourcing platforms? Does the list of URLs with high traffic, and available advertising space for speakers of a particular language (i.e. Persian), resemble that of the most visited websites in the country in question (Iran)? I conclude that separating the web and the lists of URLs may be beneficial, as a national blogosphere may have different characteristics than a crowd-sourced web.

The overall rationale for studying a national web not only implies a critique of the web as a placeless and universalized space, which is an issue concerning digital culture more widely. Location-awareness of digital media has generated the conditions of possibility for the jurisdictional to get hold of the web and more specifically through Internet censorship.[72] It allows for further development of analyses of relationships between web metrics and ground indicators. Hence the objective of the study moves towards the social research spectrum in its aim to understand the significance of national web space as one that is moving beyond the study of digital culture only. Engaging with the repurposing debate in terms of how the social is embedded into the medium as well as prescribed by it. Here I focus specifically on the relationship between metrics and measures of online device cultures and social life on the ground. I discuss metrics to analyze the liveliness of a national web, such as its responsiveness, accessibility and freshness.

[72] See for example work by Goldsmith & Wu (2006) in describing the foundational Yahoo! court case that led to the possibility to technically implement location-awareness. For Internet censorship research see for example (Deibert et al. 2008; Deibert et al. 2010).

The special case of Iran

The case study concerns Iran.[73] It is in many respects a special case, not least because the term national web itself may be interpreted as the separate Internet-like infrastructure conceived there (Rhoads and Fassihi 2011). It is also a special case in view of the scale and scope of state Internet censorship, which goes hand in hand with the repression and silencing of voices critical of the regime.[74] In other words, the Iranian web is experienced differently inside Iran than it is outside Iran. It also seems to be described in a different way, whether the author is outside or inside Iran. As a consequence many Iranians, whether site visitors or authors, whether inside or outside the country, must cope with censorship. Inside the country, coping could mean being frustrated and waiting for a friend or relative to bring news about a VPN, proxies, Google Reader and other means to circumvent blockages. Dealing with digital persecution is another matter, which is not covered in any detail here. For example, one may be warned or pursued by the Iranian cyber army (Deibert and Rohozinski 2010). One copes, or protects oneself, through the tactical selection of one piece of software or platform over another, based on which safeguards and forms of anonymity are selected, such as the easy choice of a new email address as a login in Wordpress.com and the capacity to change usernames in a Friendfeed account.[75]

Besides the reasons why Iran is a special case, certain general metrics such as site responsiveness and freshness may be put to good use when studying countries like Iran. For example, sites blocked by the state, yet still responding and updated, point to a reading audience, both outside and inside Iran. Widespread censorship circumvention may be noticeable, as is reported here. The retention of the separate webs in this sampling procedure is particularly beneficial here. The Iranian blogosphere, or the Iranian bloggers read through Google Reader and indexed and recommended by Likekhor, are squarely blocked by the state, yet they remain blogging.

[73] This case was studied in collaboration with Richard Rogers, Erik Borra, Sabine Niederer, Ebby Sharifi and the Iran Media Program at the Annenberg School, University of Pennsylvania in 2011, and with Cameran Ashraf, Bronwen Robertson, Leva Zand and Niaz Zarrinbakhsh who participated in the 2011 Digital Methods Summer School at the University of Amsterdam.

[74] Although the Iranian government continues to be known for its aggressive governance of its web (Shaheed 2014), the case in this study specifically reports on the situation of the Iranian web in the summer of 2011.

[75] Here the question of ethics also arises, where the researcher should be aware of the specific aims, purposes and objectives that are operationalized in the research (Milan 2014; Crawford, Gray, and Miltner 2014). Especially when dealing with sensitive data from users who do not want to be exposed, the researcher should consider the potential implications of pursuing avenues of inquiry where certain operationalizations might render these practices visible.

The rise of the national web marks a shift in the jurisdictional to get hold of the web.[76] Although often not recognized as such, national webs are routinely created. Developed geo-location technology facilitated national webs, serving national (or language) versions of digital media (such as Google, like Google.nl for the Netherlands) together with locally targeted advertisements and information in compliance with national laws (Goldsmith and Wu 2006; E. Schmidt 2009). At the forefront of the rise of the national and the dwindling of placelessness is the search engine whose mission statement is universal access—see Chapter 5 for the signals and algorithm updates enabling this (Google 2011b). Eric Schmidt, Google's former chief executive officer, explained that Google.com indexes information that is legal in the United States, and illegal in other countries. Google asserts that a result on Google.com is essentially controlled by Google U.S. and under jurisdiction of U.S. law. Google therefore offers local search engines, compliant with local laws. One of the earliest and most common examples used by Google executives (and by the search engine industry more in general) is that pro-Nazi content is illegal in Germany (and France), and Google omits these websites in their local domain search engines, Google.de and Google.fr (E. Schmidt 2009; Whetstone 2007). Google also abides by national youth protection laws, for instance in Korea, by enabling Safe Search by default. In such cases, Google's results page states the number of returns removed for legal reasons (Whetstone 2010). Google.cn was the best-known as well as most controversial instance of localization, whereby Google's Chinese engine drastically filtered results. In 2010 Google changed course by redirecting Google.cn (China) to Google.com.hk (Hong Kong), where the company says it does not filter, according to the company (Drummond 2010a; Drummond 2010b). The most recent example is the 'right to be forgotten' or the 'right to delist', which was enforced by the Court of Justice of the European Union (CJEU) in 2014, according to which Europeans may request search engines to delist URLs from the results based on searches for a person's name. Google has decided to enforce these locally in EU domain Googles only (Fleischer 2015).

Although the national webs are increasingly routinely created by digital media, the way to identify the location of an otherwise 'placeless' website, and to demarcate a national web space, is by no means unambiguous. How this is pursued highly depends on the assumptions, objectives and intentions that are invested in creating a national web. Another area that defines national web spaces is for instance library science, where national webs are created by national libraries and other archiving projects, which, given funding and legal constraints, also considered how to define such a web (Arvidson and Lettenström 1998; Arms 2001; Abiteboul et al. 2002; Koerbin 2003). Two notable implications of funding and legal restrictions are, first, that web archives

[76] This stands in stark contrast to the "Declaration of independence of cyberspace" by John Perry Barlow where he states that the governments have no sovereignty online (1996).

are created nationally because they are funded by the state, and second, that there is an enormous difference between the scales and scopes of the various national web archives. I am particularly interested in defining national websites, and in archivists' definitions of national webs and national websites and their implications for national web capture. For example, in previous work I found that the National Library of the Netherlands, following similar definitions of a national website from archiving projects in other European countries, defines a website as Dutch if it meets one or more of the following criteria:

1. Dutch language, and registered in the Netherlands;

2. Any language, and registered in the Netherlands;

3. Dutch language, registered outside the Netherlands; or

4. Any language, registered outside the Netherlands, with subject matter related to the Netherlands (National Library of the Netherlands 2009).

The above scheme has consequences for the collection of these websites. Ultimately, the National Library's approach could be described as editorial; especially websites related to Dutch issues, and websites in Dutch but located outside the Netherlands are particular challenging for automation, and working at scale. As a research practice, it would be impossible to automate the detection and capturing of sites that have Dutch-related subject matters (in any language and from anywhere); a list of them would probably be created, before routinely capturing them over time. The editorial, handpicking approach may be contrasted with more administrative collection practices, relying on a registrar, the national Network Information Center (NIC) responsible for registering domains, and obtaining complete lists.

Another area that is interested in demarcating and collecting the national web can be found in web characterization studies, reviewed by Baeza-Yates in 2007 and discussed at the outset. For this area the national domain (known as the country code top-level domain, or ccTLD) is the organizing entity, and one may be able to obtain a list of sites with that domain; however, in practice many countries use URLs outside their national domains, such as .com, .net and .org. As explained later, sites with the .ir ccTLD in fact may not be the preferred starting points for demarcating a national Iranian web. Another approach might be worth considering. In the data of this study the percentage of blocked .ir sites is indeed very low, compared to .com's, for example .ir seems therefore to have different characteristics than other sites authored and/or read by Iranians.

I suggest to distinguish between the definitions of a national web which are 'principled' and those that are 'adaptive', taking into account the (volatility of) media operations and their devices cultures. A principled definition is an a priori definition of what constitutes a national web and a national website, such as the librarians' above. Device cultures cover the webs formed by digital media collecting and analyzing user data and outputting leading sites of a country and/or language. Each of these digital media has an objective and interest in creating a local web informing the decisions on how it is implemented. Earlier the consequences of demarcating a national web on the basis of the formalist properties of national websites' content worth archiving were mentioned. A collection at any scale hence becomes difficult.

Having considered and discussed the 'principled' approach from the Dutch Library in defining national websites and webs, another course of action was chosen to analyze the outputs of devices, to which I will return in more detail. Methodologically, the starting point are not a priori definitions of what constitutes an Iranian website, or the Iranian web. Rather, as I will explain and defend, the URL recommendations made by dominant web devices and platforms are relied upon, which through different algorithms and logics are deemed relevant for a specific country and/or language.

The contribution to national web studies informs the literature on national web characterization, as discussed before, by developing ways to demarcate the national web and how to study it. It also contributes to censorship research, or more specifically the understudied area of self-censorship research and censorship circumvention (Canada Centre for Global Security Studies and Citizen Lab 2011). The overall approach is not only conceptual but also empirical; properties of national device cultures are identified as indicators of conditions on the ground. Such properties could include how responsive content from a specific digital medium is at any given time, and how accessible. Are responsive sites also fresh, or recently updated? Are blocked sites blocked still responsive and fresh? I am also interested in more than the technical web data sets, and how they may be repurposed for digital research. As touched upon already, in Iran in particular the websites' content is carefully monitored by the state, which is an issue that troubles digital culture and social life much more generally; websites may be blocked and website authors may be prosecuted. In the following, an approach to sample and demarcate a national web is proposed, in order to study its current conditions, including an analysis of the liveliness of the content and the extent to which that content is censored.

Demarcating the Iranian web through its device cultures

The objective of this research is to demarcate a nominal Iranian web, and analyze its condition, thereby providing indications of its liveliness in times of censorship. Here the Iranian web is demarcated through multiple, dominant digital approaches for indexing and ordering that provide a local web and privilege language, location and audience, broadly speaking. In July 2011, my colleagues and I found that the web offered by three crowd-sourcing platforms aimed at an Iranian audience differs from the web as the result of a marketing tool for Persian-language advertisers, a traffic aggregator of users in Iran, and a search engine delivering .ir sites as well as other top-level domain sites from the 'region'. They all claim, however, that they provide the Iranian web in some general or specific sense. Regarding the decision to retain the resulting data sets separately or to triangulate them, I will now discuss the outputs of each 'Iranian web' returned by different underlying principles and algorithms of the different media and how user engagement is quantified to recommend URLs. In the end I chose to write about Iranian webs in plural, and discuss each web's characteristics. I thereby addressed an issue the analyst faces when formulating where to start collecting URLs, be it in terms of compiling seed URLs to crawl, stringing together keywords and operators to form a query, consulting lists of top blogs by inlink count, top URLs by rating or top websites by hit count, etc. The outputs of the well-known aggregators of Iranian or Persian-language websites were chosen for analysis, hence not choosing one starting point, but retaining them all (or at least a number of significant ones).

Ultimately, I would like to defend a method to demarcate a national web (or 'webs') that is sensitive to the various ways web space is entered by belonging to particular device cultures, largely equated with engine and platform operations, affordances and user engagement. Generally, I introduce national web demarcation methods that repurpose digital media that are not only location-aware, but also capture device cultures. In short, I am interested in capturing national device cultures. In the analyses, I wish to chart language and other formal features present in each Iranian web. More conceptually, in this particular approach to national web studies, the liveliness of these webs is also discussed, in the sense of being online and active, or whether they are broken, in the sense of unresponsive. Additionally, I am interested in the extent to which each of these webs is censored or filtered by the state, and whether there is a relationship between responsive websites and filtered websites. In order to pursue the question whether censorship kills content, which colleagues and I described in a previous (and preliminary) project on the Tunisian web (prior to the 'Arab Spring' of 2011), we developed means to analyze if content on the Iranian web remains lively and fresh.[77] In the following, the indexing and ordering mechanisms of the web platforms and devices relevant to the Iranian space are described. The data sampled from these platforms and engines are employed to characterize the types of national webs on offer.

Device cultures and how websites are valued and ranked

Device cultures are thus understood as the specific cultures of use and social practices that take place on platforms or are indexed by search engines. Device cultures thus closely connect with the affordances of the medium and the ways in which a medium preformats and quantifies user engagement. The notion of 'grammars of action' (Agre 1994) is useful here as it underlines how engagement is inscribed into digital media and how they recursively solicit user participation in content production and evaluation. It is however good to note that there is a difference between platforms and search engines in terms of grammars of action. Whereas platforms inscribe specific grammars into the platform facilitating capture and processing of the activity taking place on the platform, as well as prescribing certain grammars outside the platform through APIs. A search engine like Google Web also prescribes grammar in its search bar, but mainly enforces its grammar, capture, quantification and processing quite differently as the search engine captures social data with a specific grammar from the web (for example hyperlinks), and subsequently quantifies and processes them in their ranking algorithm to derive a ranked list of results. The grammars of action in a search engine like Google are thus not inscribed beforehand, but rather prescribed through various communication channels such as Webmaster Help, blog posts and news items where advise about how to rank well in the search engine is communicated (for example, hyperlinks from spammy neighborhoods influence the ranking negatively). In addition, the notion of device cultures is closely connected to the understanding of how information is ordered and organized online making use of other signals, such as the locative signals that are central in the current selection of digital media.

The DoubleClick Ad Planner by Google—referred to here as Google Ad Planner—ranks sites by audience for the purposes of advertisers. Whilst 'Iran' is not among the countries listed (which is probably due to a combination of the lack of a .ir local domain Google and U.S. economic sanctions against Iran), Persian-speaking is one of the site type categories in the available audience analytics. One Iranian web could therefore

[77] This project was carried out together with Sami Ben Gharbia, Fieke Jansen and Marijn de Vries Hoogerwerff.

comprise those sites that reach a Persian-speaking audience, as collected and ranked by Google Ad Planner. With the available options 1500 unique hosts for a Persian-speaking audience were collected from Google Ad Planner.

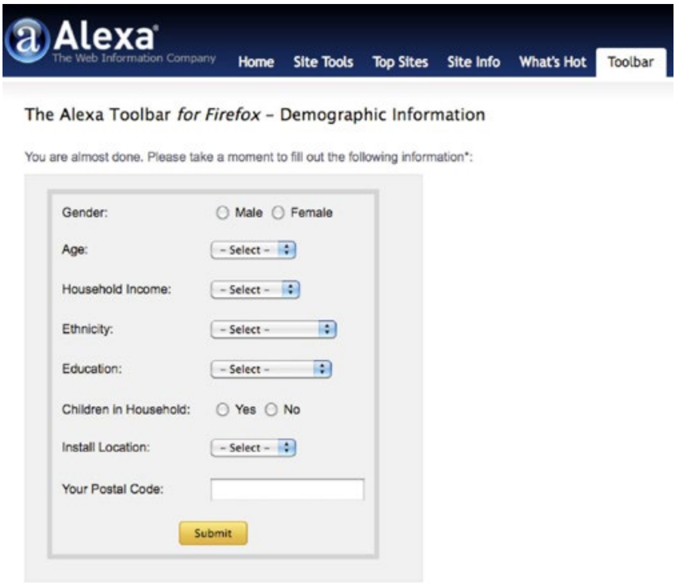
Google Web Search's key algorithms still is PageRank, which values links and a more bibliographic or scientometric manner of thinking (citation or link-counting). Although the algorithms nowadays contain more than link-counting only, as was elaborated in Chapter 5. Searching Google for .ir sites (including .ir's second level domains) as well as Iranian sites in generic top level domains in Google's regional search yielded some 3500 hosts.

Alexa, like other companies offering browser toolbars, collects user location data such as a postal code upon registration, and once the toolbar is installed, tracks websites visited by the user (see Figure 43). It thereby keeps records of the most visited sites by user location. Alexa yielded a list of the top 500 sites visited by users in Iran.

Crowd-sourced sites such as the best-known (Balatarin) and its emulators (Donbaleh and Sabzlink) require user registration before the user can suggest a link, which is then voted upon by other registered users. Those URLs with most votes rise to the top. For this study approximately 1100 hosts from Balatarin, 2850 from Donbaleh and 2750 from Sabzlink were collected. In the following analyses the two crowd-sourcing platforms Donbaleh and Sabzlink are grouped, for they share their device culture (crowd-sourcing). Together they yielded 4579 unique hosts. The Balatarin platform is treated separately because of its status as highly significant Iranian website. Launched in 2006, Balatarin is considered the first Web 2.0 site in Persian, and was recognized as one of the most popular Persian websites in 2007 and 2008 (Wikipedia 2011). It was also pivotal for the Green Movement in the opposition before and after the Iranian presidential elections in 2009 (Iran Media Program, 2010).

The introduction of the like button and other social counters in the social web led to what one may term the 'like economy,' which values content based on social button activity (Gerlitz and Helmond 2013). Likekhor, as the name suggests, ranks websites by likes; the likes are tallied from Google Reader users who have registered with Likekhor. Google Reader, or Gooder, is particularly interesting because it allows to read the content of websites otherwise filtered by the state, thus effectively acting as a proxy to filtered websites. Likekhor focuses on blogs, indicating a relationship between Google Reader users and bloggers, or blog readers. A list of 2600 hosts was extracted from Likekhor collected from a page listing all blogs on Likekhor.

In this way in July 2011 just over 10,000 hosts were collected through platforms and devices significant to Iranian users (Google Reader, Google Web Search and the



The Alexa Toolbar *for Firefox* – Demographic Information

You are almost done. Please take a moment to fill out the following information*:

Gender: ○ Male ○ Female

Age: - Select -

Household Income: - Select -

Ethnicity: - Select -

Education: - Select -

Children in Household: ○ Yes ○ No

Install Location: - Select -

Your Postal Code:

Figure 43: Alexa toolbar installation and registration process, with a field for the user's postal code. Partial screenshot, 2011. Source: *Alexa n.d.*

crowd-sourcing platforms) and through two providing ranked lists of Iranian or Persian-speaking sites (Alexa and Google Ad Planner) on the basis of data collected from users located in Iran (Alexa) or from Persian-writing users (Google Ad Planner). In the following these Iranian webs are characterized, triangulated and compared individually and against the triangulated Iranian web. The findings, as alluded to above, benefit from treating the webs separate for the analysis. Few websites recur across the variously sourced Iranian webs.

Analyzing the characteristics of device cultures

In web characterization studies one of the main difficulties repeatedly discussed is how to obtain a representative sample of a national web or other web types. According to Baeza-Yates and colleagues, the three common types of sampling techniques used in web characterization studies are 'complete crawls of a single web site, random samples from the whole web, and large samples from specific communities' (2007, 1). For national webs, which the authors consider to be specific communities, the list consists of websites with the same ccTLD. For many national webs, however, this restriction would be too limited, especially for countries where generic top-level domain usage is prevalent. Our approach sought to retain the .com's, .org's, .net's, etc. when deemed relevant for Iranians and Persian-speakers by the devices and platforms we relied upon.

I would like to add a fourth type, multiple aggregator site scraping or more conceptually said, device cultures to the sampling techniques described above. In any case, Google Ad Planner, Alexa, Google Web Search, Likekhor (Google Reader) as well as the crowd-sourcing platforms (Donbaleh, Sabzlink and Balatarin) either through query results or (dynamically generated) listings yield websites relevant for Iranians and Persian speakers. In the case at hand, with the exception of the searcher's web (gained through .ir and generic TLD queries in Google's region search), the percentages of .ir sites among the significant hosts outputted by the devices were relatively low (see Table 3). The crowd-sourced web references yielded the fewest .ir sites at just over 10 percent, whilst both Google Ad Planner's web as well as Alexa's web yielded most at about 25 percent. As noted earlier, the .ir sites in the overall collection of URLs were much less likely to be blocked than the .com sites. 80 percent of the websites tested and found blocked from inside Iran were .com, followed by .net with 6 percent and .org with 4 percent. The ccTLD .ir contained 3 percent of all censored hosts.

Percentage	Iranian Web	Absolute numbers
25%	Alexa	126 of 496 hosts
24%	Google Ad Planner	370 of 1525 hosts
16%	Likekhor/Google Reader	397 of 2541 hosts
12%	Donbaleh/Sabzlink	535 of 4579 hosts
11%	Balatarin	116 of 1102 hosts

Table 3: Percentage of .ir sites in top websites collected from Alexa, Google Ad Planner, Likekhor, Donbaleh/Sabzlink and Balatarin, July 2011. Source: *Rogers et al. 2011.*

Having reviewed how samples are generally made, Baeza-Yates and colleagues compared the ten national web studies performed so far, in order to arrive at a core set of measures shared across many of them (see Table 4). Our study's particular point of departure in characterizing the Iranian web (or webs) benefits from the metrics available. Referring to the metrics in Table 4, in the category of content our project shares interest in language, page age and domain analysis (albeit top-level), and in the category of technology it relies on HTTP response codes. The codes yield what is referred to as 'responsiveness,' which is considered a basic metric, together with page age, and the measure of freshness.

Content	**Link**	**Technology**
Language	Degree	URL length
Page size	Ranking	HTTP response code
Page age	Web structure	Media and document formats
Pages per site		Image formats
Sites & pages per domain		Sites that cannot be crawled correctly
Second-level domain		Web server software
		Programming languages for dynamic pages

Table 4: Metrics commonly used in national web characterization studies, 2011. Source: *Baeza-Yates, Castillo, and Efthimiadis 2007*.

Languages

One basic metric seeks to measure the composition of languages in the Iranian web (see Figure 44). Persian is of course the official language in Iran; the Unicode system incorporated Persian script in 2001, and it can be detected (Amir-Ebrahimi 2008). For the language detection of websites we built a custom tool that makes use of the alchemyAPI, and is able to detect Persian as well as other, but not all languages spoken in Iran, as I will address shortly. In a second step, the results were checked manually. Two out of three sites in the Iranian web, in total, were in Persian, English being second with one out of five. The proportions of the use of Persian in the various webs are interesting. The results show that the bloggers' space, Likekhor, ranks highest with 91 percent of the sources in Persian, followed by Alexa's Iran-based surfer's web with 83 percent and the crowd-sourced web with 73 percent. At the bottom are Google Ad Planner's web with 62 percent, and Google Web Search with 52 percent. Balatarin, the special case, has 75 percent in Persian. There is a significant difference between the webs, including, notably, a Persian-dominated web in the Likekhor/Google Reader space.

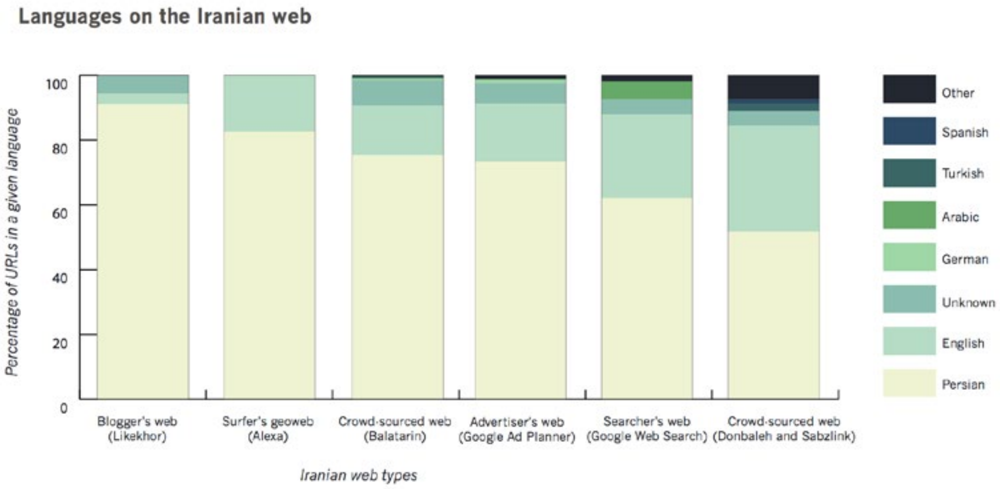


Figure 44: The distribution of languages on the Iranian webs. Likekhor, Alexa, Balatarin, Google Ad Planner, Google Web, Donbaleh and Sabzlink were queried to retrieve (ranked) lists of relevant URLs. As platforms such as Alexa only provide a ranked list of hosts, all URLs were chopped to their (sub-)domain part, in order to facilitate comparison. The data were collected on July 7, 2011 and resulted in a list of URLs per web device. To automatically detect the different languages used in the Iranian webs, a custom tool was used that makes use of *AlchemyAPI*. The graph shows the percentage of URLs in a given language, per web device. The languages are color-coded. Likekhor's web is mainly in Persian, while the Donbaleh and Sabzlink's web has most diverse languages. Visualization created in Adobe Illustrator, 2011. Source: *Rogers et al. 2011*.

Here I would like to briefly discuss the kinds of webs that would be captured and analyzed if the Iranian web or an Iranian website were to be defined a priori, which I would I would endeavor according to a principled definition, a subject matter raised earlier with respect to the web archivist's formal conditions of a national website (in the Dutch example). The Likekhor/Google Reader web and to a slightly lesser extent Alexa's web (based on surfers in Iran) give the impression that an Iranian web is Persian-speaking only, although we would still have to reckon with an average of over 10 percent of non-Persian websites. The Iranian webs with higher percentages of non-Persian sites are Google Ad Planner's web as well as local domain Google's web. Google Ad Planner is accessed by Persian speakers as detected by the signals Google compiles on its users and the content indexed, and these webs have far higher percentages of non-Persian sites, especially English. No attempts were made, however, to investigate whether these sites are authored by Iranians, or concern Iranian affairs, in whatever way that may be defined. Another web could be conceived a priori, which also would have implications on the method used to construct the object of study. Being all-inclusive in terms of the languages spoken in Iran (Armenian, Assyrian Neo-Aramaic, Azeri, Kurdish, Lori, Balochi, Gilaki, Mazandarani, Arabic and

Turkmen) also has consequences for the capturing techniques; of the secondary languages spoken in Iran, the language detection tool employed in this study only detects Armenian, Arabic and Azeri. Specialists' link lists would have to be sued, although we did not pursue the matter any further.

Responsiveness

To analyze the responsiveness of the Iranian webs the HTTP response status codes were retrieved with a custom built tool. The lists of hosts from the previously collected Iranian webs were the inputs. Analyzing the results returned by the response code tool, eight commonly returned response codes were found in the Iranian web spaces (see Figure 45). The 400 class of status codes indicates that the client has erred; '404 not found' is considered the strongest indication of unresponsiveness. '400 bad request' means that there was an error in the syntax, '403 forbidden' indicates that the server is refusing to respond. '404 not found' means that the content is no longer available. Commonly returned response codes besides the '200 OK' status, are two redirecting response codes: '301 moved permanently' and '302 found.' Redirecting is not necessarily an indication of unresponsiveness; it can have a range of reasons, including forwarding multiple domain names to the same location, redirecting short aliases to longer URLs, moving a site to a new domain, or it may be an indication of a parked website. However, redirects may also indicate 'soft 404' messages to hide broken links (Bar-Yossef et al. 2004). In the current study both 301 and 302 were followed if a location header was returned, which mostly resolved in 200 and 404 response codes. '0 connection problem' indicates that the tool was unable to connect to the server; the server may no longer exist, or it means that our tool timed out.

The first findings of this portion of the study indicate that the Iranian web(s) were relatively lively overall. The crowd-sourcing webs of Donbaleh/Sabzlink and Balatarin resolved respectively 92 and 94 percent of the sites. The Google Ad Planner space, followed by the Likekhor space, delivered through Likekhor to Google Reader users, was the liveliest, with 96 and 95 percent of the websites resolving. The vibrancy of the (Persian-language) Google Ad Planner space, Likekhor/Google Reader as well as the crowd-sourced webs was thus established.

Internet censorship

Arguably, digital media are among the most well-informed censorship monitoring instruments. Search engines and platforms receive requests to delete content—either specific URLs and queries or more general instructions—thereby creating an ongoing

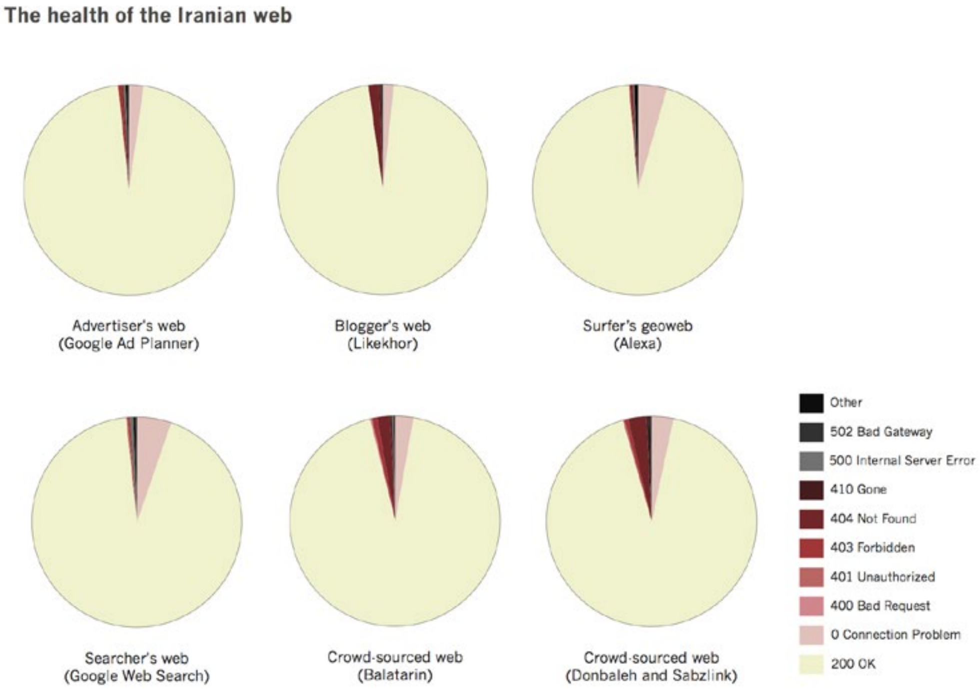
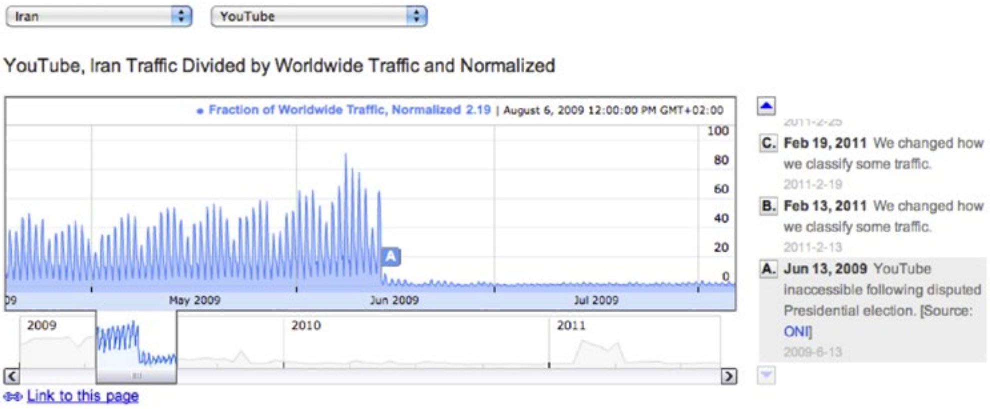


Figure 45: The liveliness of the Iranian webs measured by HTTP response codes retrieved from the Netherlands. The graph shows color codes indicating the response codes and is subdivided in pie charts per web device. Data was collected in July and August 2011. The crowd-sourced web had most blocked pages. Visualization created in Adobe Illustrator, 2011. Source: *Rogers et al. 2011*.

blacklist as well as a censorship index. For example, in line with the Chinese government's censorship instructions (prior to the redirect to .hk), Google engineers were reported to 'set up a computer inside China and programmed it to try to access websites outside the country, one after another. If a site was blocked by the firewall, it meant the government regarded it as illicit—so it became part of Google's blacklist' (Thompson, 2006). In the case of the Iranian web, one of the most aggressively censored webs in the world, the government was not reported to request removal (Open Net Initiative 2009; Google 2011a). However, the graph in Figure 46 shows how Iranian traffic to YouTube increased in the run-up to the presidential elections in June 2009, before coming to an almost complete standstill one day after. The interesting question in this study is to what extent blocking important sites had implications for the liveliness of the Iranian webs. Next, by means of proxies it was checked if the collected Iranian webs were available inside Iran. These findings were compared with the basic health measures, responsiveness and freshness. As mentioned above, one of the


Iran
YouTube
YouTube, Iran Traffic Divided by Worldwide Traffic and Normalized
• Fraction of Worldwide Traffic, Normalized 2.19 | August 6, 2009 12:00:00 PM GMT+02:00
100
80
60
40
20
0
A
09
May 2009
Jun 2009
Jul 2009
2009
2010
2011
Link to this page

2011-2-25

C. Feb 19, 2011 We changed how we classify some traffic.
2011-2-19

B. Feb 13, 2011 We changed how we classify some traffic.
2011-2-13

A. Jun 13, 2009 YouTube inaccessible following disputed Presidential election. [Source: ONI]
2009-6-13

Figure 46: Iranian traffic to YouTube comes to a standstill after the 2009 presidential elections, August 25, 2011. Source: *Google 2011*.

more remarkable findings were that a large portion of the Iranian blogs was blocked, yet continued to respond, and was rather fresh.

The Censorship Explorer tool lists (fresh) proxies by country, and may be used to check for censored websites. The tool returns website response codes or loads the actual websites in the browser, as if you were in the chosen country in question. As a starting point in the censorship research procedure, website responsiveness is often checked in a country not known to censor (Iranian) websites (in this case, the Netherlands). Subsequently, lists of hosts are run through proxies in Iran, and the response codes logged. Iranian servers typically return the 403 forbidden response code, which is a strong indication of a site being blocked (Noman 2008). Response code checks through proxies may give an indication of specific types of Internet censorship: URL and IP blocking, which includes censorship techniques such as TCP/IP header filtering, TCP/IP content filtering and HTTP proxy filtering (Murdoch and Anderson 2008). (Other known filtering techniques are more accurately detected by other means, including DNS tampering and partial content filtering.) Multiple proxies allow the researcher to triangulate proxy results and increase the trustworthiness of the results. For example, '0 Connection Problem' may be a proxy problem, but may just as well be an RST package returned by the censors, resetting and effectively dropping the connection (Villeneuve 2006). Comparing multiple proxies can confirm that it is not a proxy problem. In this study twelve proxies were used, hosted in six different cities in Iran and operated by various owners, including Sharif University of Technology and the popular Internet service provider Pars Digital. Concern has been voiced that it is 'false to consider Internet filtering as an homogeneous phenomenon across a country,' considering that both the implementation and user experience of censorship may vary by city, ISP, or even by computer (Wright, De Souza, and Brown 2011, 5). Taking this concern into account, we selected proxies from different cities and ISPs, and subsequently considered the response code returned by the majority. Table 5 details which proxies were used for this research.

IP address and port number vf proxy	ISP and location of proxy
217.219.115.133:80	ITC, Tehran, Esfahan
91.98.137.196:80	Sharif University Of Technology, Sharif, Khuzestan
78.39.55.11:3128	ITC, Fars, Shiraz
91.98.137.196:3128	Pars Digital, Tehran, Esfahan
80.191.120.129:3128	ITC, Tehran
213.217.43.82:8080	Pars Digital, Pars, Tehran
217.219.115.137:80	ITC, Tehran, Esfahan
217.219.97.11:3128	ITC, Shiraz, Fars
80.191.122.11:3128	ITC, Shiraz, Fars
80.191.227.243:3128	ITC, Ahwaz, Khuzestan
188.136.241.2:3128	Ariana Gostar Spadana, Esfahan, Esfahan
188.136.156.116:3128	Ariana Gostar Spadana, Gostar, Hamadan

Table 5: Details of the proxies used to test for censorship in Iran. URLs were queried through the various proxies between August 22, 2011 and September 8, 2011. Proxies retrieved from the Censorship Explorer tool.

The results in Figure 47 show that approximately 6 percent of Alexa's (29 out of 497) and 16 percent of Google Ad Planner's web (241 out of 1525 hosts) was blocked. The crowd-sourced web had just over 50 percent of the web blocked, with 2411 of 4579 hosts. Balatarin was the most aggressively censored Iranian web space with 58 percent blocked, or 639 of 1102 hosts, followed by the other two crowd-sourcing platforms—Donbaleh and Sabzlink—with more than half of the hosts blocked. Likekhor/Google Reader's web, which in the research work thus far represented the Iranian blogosphere, had 1137 of 2541 sites returning the 403 forbidden code, meaning that 45 percent of it was blocked. As discussed above, Likekhor/Google Reader's web is largely in the Persian language, and was one of the most responsive of all the webs studied, with 95 percent of the sites returning 200 OK response codes. This proved

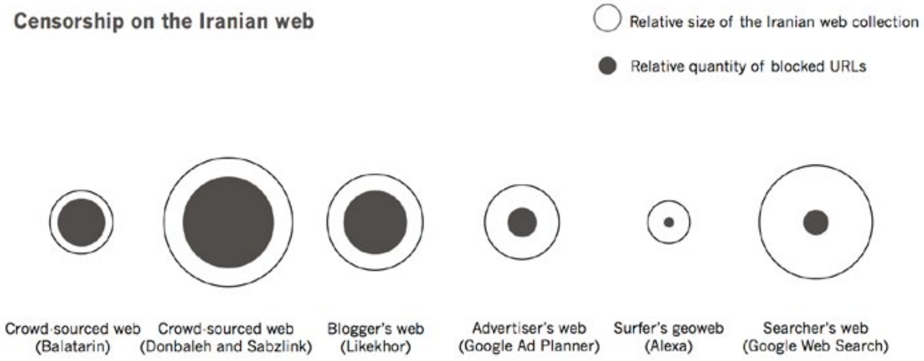


Figure 47: Censorship in the Iranian webs, August 2011. In order to test for censorship all the URLs retrieved from the web devices were run through proxies located in Iran with the use of the Censorship Explorer tool. To obtain a list of proxies in Iran a web search for 'free proxy lists' was done and it was made sure that a diverse range of proxies, located in different cities and in the IP-range of different ISPs, was obtained. The graph shows the relative quantity of blocked URLs per web device. The crowd-sourced web and the Blogger's web have most blocked pages. Visualization created in Adobe Illustrator, 2011. Source: *Rogers et al. 2011.*

the Google Reader usage to be a vibrant censorship circumvention culture. Google Reader uses RSS to redistribute content through the web. RSS is a format that makes use of the HTTP protocol to syndicate content, which makes it more challenging to censor (Zittrain and Palfrey 2008) than the URL of a website because RSS and similar technologies are open system in that they allows external tools and sites to read and distribute content from it. This study appeared to render censorship circumvention visible at a large scale; the least one can say is that blocked websites were still online. Of the webs checked for filtering, the crowd-sourced sites as well as the Likekhor listing are most often blocked, which raised the question not only of the substance of those spaces (I will treat Balatarin's below), but also whether the platforms as URL lists are convenient for monitoring. Whilst many sites were blocked but still responsive, other signs of liveliness drew our attention. Were they fresh? If the sites are blocked, yet responsive and fresh, there is a strong indication that censorship is not effective (to date).

Freshness

After identifying the spaces of particular interest (crowd-sourced as well as the blogger's webs), and after finding them highly responsive as well as heavily blocked, the question was whether censorship kills content. Or, despite having their sites censored, did the bloggers keep on blogging, and did the crowd keep posting? Can readers be expected to routinely circumvent censorship, and can content continue to be recommended, commented on, et cetera? Apart from being responsive (nearly all websites were found online), it is necessary to know whether they are active. Is the content on the websites fresh? The study at hand researched a subset of the webs—the blocked sites in the crowd-sourced and the blogger's webs. To determine how fresh these sites were, we asked the Google feed API for each host (per list) whether it had a feed (for example RSS or atom). If so, the feed was parsed with the Python Universal Feed Parser library and the date of the latest post extracted. Overall, 63 percent (5147 of the 8222) of the three webs had feeds. Of the blocked sites in these webs, 71 percent (2986 of the 4189) had a feed. For Balatarin, the percentage of blocked sites with a feed was 79% (504 of 639 blocked hosts), for Donbaleh/Sabzlink 68% (1630 of 2413) and for Likekhor 75% (852 of 1137).

What constitutes a fresh site? One may turn to blog search engines for advice about staleness and freshness. In an FAQ about blog quality guidelines, Technorati stated that they 'only indexed 30 days' content, so anything older than that did not appear on Technorati' (Technorati 2011). Similarly, search engine and analytics system for blogs—Blogpulse—takes 30 days as a measure of fresh content: 'A blog's rank is based on a moving average of its citation counts over the past 30 days' (Blogpulse, 2011). Thus, freshness here was defined as having at least one post published via a feed in the last month, counted from the moment we last checked for blockage. Could these sites expected to be fresh? To draw the findings into stark relief, it is interesting to note that the well-known survey conducted by Technorati in 2008 found that about 7 million of the 133 million blogs it followed had been updated in the past four months. *The New York Times* wrote that the finding implied that '95 percent of blogs [were] essentially abandoned, left to lie fallow on the web, where they become public remnants of a dream—or at least an ambition—unfulfilled' (Quenqua 2009). In this study 65 percent of the sites overall were found to be fresh. In the crowd-sourcing platform Balatarin 78 percent of the blocked hosts with a feed (395 of 504 hosts) were fresh, and in the crowd-sourcing web organized by Donbaleh and Sabzlink 56 percent of the blocked hosts with a feed (915 of 1630 hosts) were fresh. For the Likekhor list, 61 percent—or 525 hosts—had a post dated a month before they were tested and found blocked. The results confirm that there is hardly a general indication that censorship kills content on the Iranian web. On the contrary, the most severely censored Iranian webs were both responsive and rather fresh.

Conclusion

The research reported on in this chapter is a contribution to demarcating the national web, and an approach to study it. Additionally, in this special case of Iran a contribution to censorship research is made by developing ways to empirically inquire into the effectiveness of censorship and by studying censorship circumvention.

First, a methodological plea for capturing and analyzing the diversity of national web spaces, or webs is made. Rather than predefining national websites, and thereby national webs, a principled approach of formal properties (for instance, all websites with ccTLD .ir, all websites in Persian with Iran-related content, or websites with authors inside Iran), lead to the conclusion that such approaches can often not be operationalized or automated. Instead I proposed to use so called 'device cultures,' and in particular the Iranian web spaces they provide. Device cultures more specifically are defined as the interaction between user and platform or engine, the data routinely collected, how they are analyzed, and ultimately the resulting recommendations. National webs are demarcated through location-aware devices; location or language is added as a value sifting out URLs relevant to Iranians and Persian-speakers. When examining the data sets and performing a top-level domain analysis of the sample, the majority of the collected hosts from the various Iranian webs were found to be .com websites, not .ir, which expanded the scope of national domain characterization studies, and introduced a method of data collection for broader national web studies.

Second, when building on national web characterization studies liveliness and contributing to censorship research, indicators were used to detect the effectiveness of self-censorship. In so doing this chapter further explores digital research with realtime devices and dynamic data by taking an approach and fully benefitting from the algorithmic nowness of realtime media: the research measured the liveliness of the Iranian web. The Iranian web was studied with a number of devices by retaining a device's specific culture and by comparing the liveliness across devices instead of merging the data. More specifically, liveliness was conceptualized as a series of metrics, a limited number of which was employed in this study, that is, responsiveness, page age and filtering or blockage. (In addition language detection and top-level domain analysis were performed.) The contribution of this work to Internet censorship research is three-fold. The first is the device-driven demarcation tactic using digital media as described above. The second is conceptual, in that it proposes to repurpose metrics from national web characterization for liveliness indicators. Are websites responding? Are pages fresh? The third can be generalized for countries facing state censorship. The results from the responsiveness tests are compared to the filtering ones. Are the blocked sites still responsive and fresh? The approach showed that a considerable number of blogs were blocked, yet still responsive. This indicated an audience for the content, both outside Iran and inside, which hinted at widespread Internet circumvention in a particular space: the predominantly Persian-language blogosphere authored by Likekhor and Google Reader which in tandem served as an important filter for Iranian blogs. Although heavily censored, the Iranian blogosphere as listed by Likekhor remained vibrant. This censored but active space was similar to the crowd-sourced web, organized by Balatarin. Blocked yet posting, Balatarin's recommended websites also suggested a similar finding as the one for the blogosphere: an active audience for blocked websites.

Third, the device culture approach to the national web does not only imply a critique of the web as placeless space, which is an issue that concerns digital culture more widely; it is also a means to further analyze relationships between web metrics and ground indicators. I thereby further develop the third contribution to digital research, which is how the social is inscribed into the medium and prescribed by that the medium, focusing specifically on the relationship between metrics and measures of device cultures online and social life on the ground. Certain implications of national web studies affect both current and future policies with respect to the web (and its study) and the use of web indicators for social research in general.

Conclusion: What is a good digital device?

Chapter 8

ONE OF THE KEY objectives of this dissertation is to examine the ways in which collections of data harvested from platforms and engines shape, and are shaped by, digital research. The case studies investigated in detail how understanding the operative capacities of devices benefits from an analysis of the ways in which software features, processes, and operations participate in 'doing research'. The proposed concept of research affordances turns digital research into an on-going process of assembling, reconfiguring, and aligning research questions with digital media and device cultures. Throughout the term device was not merely used to address the operative capacities of platforms and engines, but also to refer to the purposes written into these digital platforms and engines. I use the notion of device to fold object and method into each other as it enables me to talk about the performative qualities of platforms and engines and about their affordances for research. This allows me to conceptualize digital research as both material and practical, constructed and designed in certain ways, and as a procedure for assembling and organizing what it pertains to. Building on a software studies informed understanding of digital media as a productive force, I illustrated some of the ways in which digital media can be operationalized differently in very specific research scenarios. Medium research and social research, and any combination of the two, can be very obvious as well as very advanced in its configuration. Both devices and research objectives inform research processes and culminate in digital methods. The notion of research affordances is crucial for grasping the role of digital media as devices in the method.

Juxtaposing the research affordances of different media prompts reflection on how devices privilege certain research questions rather than others. The critiques to the Twitter election prediction study addressed in the introduction indicate the complex tensions underlying the use of web media and their data in digital research; data are not just 'there' (Gitelman 2013). By proposing a shift in focus from data-driven to device-driven research I have been exploring how these tensions play out on the various abstract levels introduced in this dissertation: both at the level of aims and objectives brought to the data by the researcher's questions, and at the platforms' and engines' medium-specific level. Repurposing means to operationalize research aims through device affordances and use practices; research questions are thus in dialogue with the social and cultural enacted through online media. This includes but is not limited to the role of algorithms in the steering of analytic processes and of specific features and functions to solicit and structure analysis. Focusing on the research affordances of digital media underlines that the operations of platforms and engines as analytical devices are taken seriously, serving as sources of data for digital research. Shifting focus from data to the device also allows us to start asking questions about the ethics, politics and economics embedded into our digital media. It allows us to inquire into the specific algorithms and signals that are used in the back-end, the specific functions and features that are offered in the front-end, and how they together inform the specificity of operations, and their related concerns and issues, that make up our digital media that are key to our digital culture and social life.

Digital methods has become a keyword. A Google Scholar Alert on the term regularly informs me that a range of areas use the term, either just to indicate the methods that operate digitally, or (a version of) the more conceptual use of the term as it is employed here. In this dissertation I contribute to digital methods not just by specifying that a method works digitally or on digital data, but by calling attention to how online platforms and engines may also be used as devices to do research with. A number of areas dealing with the computational or digital turn in social research more broadly engage with questions related to the quality of method and data (Markham and Baym 2008; Borgman 2009; Rogers 2013b; Liu 2015). In this concluding chapter I seek to contribute to these efforts by addressing some of the key concerns related to digital devices dealt with throughout this dissertation and by connecting these to larger debates in digital research. I juxtapose what I consider to be principled evaluative criteria and stress that I do not consider the quality of a device as an inherent property. Rather, a device becomes good *choice* if it can be productively aligned within the whole methodological configuration. Within that configuration, I have been arguing, both medium and social research can be focused on, and the quality of the device as component in the research apparatus is the interplay of its productive capacities, the research question and the research objective.

Research affordances

Central to the repurposing debate is the question whether one is studying the medium or studying the social. Although this distinction sounds elementary, the specific concerns addressed are quite complex. My contribution to the repurposing debate is that I shift the focus of the debate to the role of the digital device in the research process. This distinction between studying the social and studying the medium can always be seen as a difference in degree: it is always a bit of both, as are the findings. With change of focus towards acknowledging and evaluating the digital device as an analytical component in the research apparatus I shift the debate to the quality of the configuration and the 'sophistication' (Fuller and Goffey 2012) of the operationalization.

The key concerns, then, are how to operationalize the medium for research purposes and how to align the various components of the research apparatus? Or, how to avoid turning a blind eye to the assumptions and norms the digital medium brings to the research and instead how to employ the media operations by configuring them in relation to the research's objective? Repurposing—operationalizing research aims through medium operations—turns digital research methods often in rather complex assemblies of the different forces involved.

What makes a digital device productive, if not for social research then for medium research? To answer this question the focus is shifted towards the research affordances of online media. Both medium and social research methods can be inventive and advanced in their operationalization, whereas the configuration of research objective, digital medium and device culture can at the same time be very obvious or even incorrect. With the device perspective, the focus shifts to the relations between digital methods and the digital media it collects data from, highlighting the tensions in the research process emerging when working with digital data. With the overarching notion of 'research affordances', in the following I will return to the three key concerns within device-driven research closely connected to issues troubling digital culture and social life more generally.

Medium dependency

My first contribution relates to the fact that digital research can have different research aims: social findings versus findings about the medium. The difference between social research and medium research was exemplified in the case studies on realtime in Chapters 2 and 3. Although both engage with the research affordances of digital media making them operate in the 'now', the specific features and processes repurposed in the respective case studies focus on operationalizing them, geared towards social findings (i.e. 'lively' content) and medium findings (i.e. the constructions of realtime). In both propositions, the operative capacities of engines and platforms as devices in the research process are the key focal points. This device-driven perspective brings a perspective informed by software studies to social research, and offers medium research, that is data analysis as afforded by the software or medium under question, to software studies.

From a social research perspective digital media, when considered as sources of data, might introduce 'alien assumptions' into the research practice. This notion of 'alien assumptions' introduced in Chapter 2 connects to larger debates in digital research concerning the supposed 'messiness of digital data'. But what is messy in web data? Not all authors have similar opinions on the quality of web data (see for example Savage and Burrows 2007; Uprichard 2013; Thelwall, Prabowo, and Fairclough 2006). When data from online media is cleaned for the purposes of social research, however, built-in assumptions which do not stem from the researcher's design, but rather are 'native' to the data collected from platforms and engines, are often removed. In a short position paper Alan Liu situates the notion of 'alien assumptions' in the larger context of digital humanities and states 'that's the promise of big data' (Liu 2015), as it allows asking new types of questions. It provides an opportunity to transact between a researcher's 'familiar domain and scary alien ontologies, epistemologies, forms, sociologies, economics, and politics'; he continues that such a transaction would '"operationalize" the messiness' of data as something more than accident or inconvenience (Liu 2015; see also Moretti 2013). In device-driven research, then, what is considered messiness is not an a priori given but depends on the research objective. In contrast, the device perspective calls out to recognize and evaluate the research affordances of digital media as opportunities for new modes of digital research.

Using the medium's operations and the resulting data to study the medium itself contributes to the area of software studies by engaging software in such a way that it can be used to operationalize its own critique. This digital methods approach responds to calls made in software studies to study software 'in action' (Manovich 2012a; Goffey 2008; Sandvig et al. 2014) as well as to the call for more methods (Manovich 2013). The black-boxedness of knowledge technologies is one of the key methodological concerns in this context (Gillespie 2014; Fuller and Goffey 2012; Sandvig et al. 2014; Bucher 2012a). Software studies have creatively resorted to a variety of materials to study the often black-boxed proprietary software of dominant digital media, for example by analyzing patents regarding the origins and intentions behind specific algorithmic configurations (Rieder 2012), or tech documentation in combination with interface analysis to connect the interface to the code (Bucher 2012a). The digital methods approach advanced in this dissertation also uses these medium-specific materials and adds to it by engaging the operations and outputs of black-boxed software in the research process.

Volatility

The second contribution connects to the issue of volatility and specifically the extent to which methods can handle the volatility of digital media. Digital media often update the way in which they capture, format, and recommend information. This volatility of digital media calls for attentiveness to the specific workings of the medium when operationalizing research affordances in relation to the research aim. On the other hand it also clarifies how the connection between research affordances and modes of research may change the problem at hand—what Lury and Wakeford referred to as the need for the inventiveness of methods (2012). In digital research the volatility of digital methods thus signals the intricate relation between medium and method; in the method the research aim is negotiated through the research affordances and cultures of use of the medium (see also Marres and Gerlitz 2015). Digital research is therefore always partly medium studies and partly social studies, as it highlights the specific purposes and use practices inscribed into digital devices and the consequences of specific configurations on the findings. In this dissertation I have thus been engaging with volatility by inquiring into the relative delicacy of the research apparatus, how devices may change the problem at hand, as well as how the devices themselves may be volatile due to the continuous update cultures of digital media.

The notion of volatility is also closely linked to what may be considered the selection rigor of the digital media, or the basis on which things are included in the database. When dealing with the selection rigor of platforms and engines, work in the field of 'platform politics' (Gillespie 2010; Gehl 2011) is relevant, since platforms negotiate interests between different parties; these interests in turn shape what data points are or are not collected. The interests of advertisers, users, third party developers, policy makers, governments may all at some point be decisive in shaping or appropriating the purpose of the platform. By the same token, the databases of digital platforms and engines tend to collect broadly for strategic reasons to accommodate various interests. Some digital research disciplines mimic this collection strategy in their method, most notably those operating under the header of big data research, where more data means better data (Bollier 2010; Manovich 2012b; Lazer et al. 2009). Alongside this data-driven approach I have been making a case for the device-driven approach being a slight, yet significant, shift in focus in digital research. Here the notion of repurposing is constructive since the selection criteria and demarcation of the data set are the result of the negotiation between the research question and the specific conditions and limitations of the exhaustive collection of the platform or engine. The relation between the concept of 'research affordances' and the notion of 'volatile' as opposed to 'principled' method, as discussed in Chapter 7, highlights the difference between 'following the medium' (Rogers 2013b) to sample digital data, as opposed to collecting data following an a priori schema. Digital methods are adaptive methods; their operation is not ensured by principle but by tactic, not from the outside but embedded in the object, not by a theoretical a priori but by practice. Digital methods adapt to how devices organize data in specific ways, thus provoking a change in status of this type of epistemic trouble. In digital research, epistemic issues such as volatility trouble research in a rather immanent fashion: they affect it from the inside but may also become themselves an object of research.

Device cultures

The third issue concerns the notion of 'device cultures,' or how the medium's affordances prescribe cultures of use and social practices, as well as how the social and the cultural at the same time inscribe into the medium. I have been connecting the notion of ‘research affordances’ to the notion of ‘grammars of action’ (Agre 1994), to conceptualize how the medium prescribes cultural uses and social practices. Phil Agre argues that action is achieved through software, because many of the activities do not exist outside the grammars inscribed in it (Agre 1994). Digital devices and their research affordances thus shape the operations and activities enacted through them. In terms of the ‘grammars of action’ encoded into digital media, a general shift has increasingly turned digital media into applied services rendering them more easy to use but at the same time, by introducing more fine-grained and scripted grammars, rendering them more difficult to tinker with (Van Dijck 2013; Zittrain 2008, 104–107). The analysis of the historical Dutch blogosphere in Chapter 4 illustrates how messy, general-purpose blogging features and functions developed into closed and off-the-shelve software. Software grammars are more and more tightly pre-formatted and off-the-shelve. These features and functions as affordances imply that they increasingly pre-structure potential actions that may be assembled like building blocks by third parties, including digital researchers. Although digital media shape the conditions of possible actions, the construction of digital media and their cultural uses and social practices are mutually constitutive (Halavais 2014; Van Dijck 2011).

The notion of device culture it attentive to the grammars of action inscribed in the medium; it also takes into account how digital media and their affordances are used in practice and operationalized in the method. Device cultures can be made operational to show what is the most popular, lively, controversial, and so on, depending on the specific configuration in relation to the research objective.

Because of the tactic of repurposing the various digital media must instinctively be treated separately as they have their own logics. Chapter 7 showed how data from various digital media may be compared without losing their distinctiveness. The research reported on in the chapter repurposed the specific cultures of six digital media and exploited their distinctiveness to demarcate a national web. Not only does this device-driven perspective allow for the demarcation of the national web, it also proved to be a productive way to sample a national web and find the most relevant URLs on the basis of the ranking mechanisms and frequency dynamics of significant media. They provide a valuable starting point to research the extent to which areas of the national web are censored, and whether or not censorship leads to self-censorship measured through the decline of content production in specific areas of the web.

A guiding principle for evaluating the device is the alignment between the goal of the device and the direction of the method, which, to me, is captured by the term ‘repurposing’. Moreover, devices privilege specific questions and types of research. Repurposing is a deliberate application of the medium’s research affordances combined with how it is practiced by users, a negotiation between different forces at play. By slightly shifting the focus in the repurposing debate to the digital device, the quality of the method does not necessarily depend on questions like how much medium, how much social, but instead it is based on the quality of the configuration. Key for reliable and interesting results is the identification of the sites of negotiation when configuring research objective, digital media and device cultures in the research apparatus.

Modes of research

In the following I revisit the specific configurations deployed throughout the chapters as I go through the different modes of research afforded by the platforms and engines under study. The modes of research align device operations in the method with the research objective. The case study in Chapter 2 started by differentiating between 'liveliness' and 'liveness' in realtime research which, in brief, is the difference between measuring frequency versus measuring engagement. With the rise of the realtime web, the web as a source for social research risks being caught in an 'eternal now'. What can digital social research do with dynamic digital data? The chapter sheds light on and provides a periodization of the mundane practice of two dominant forms of data collection: scraping and calling APIs, and advances a methodological contribution to the role of scrapers and APIs in social research by highlighting the fact that they capture, format, and to a certain extent already analyze data. The case study put questions of liveness versus liveliness into relief, with as one of the main challenges to differentiate more precisely between the study of media dynamics and social dynamics.

Chapter 3 on the politics of realtime engaged with 'realtime research' as 'medium research', instead of the types of social research enabled through it. The comparative study of various online media was used to empirically gauge how they construct realtime. The case study thus presented an empirical contribution to a theoretical debate by introducing the notion of 'realtimeness', a device-specific account of the construction of realtime. Media do not operate in realtime; devices and their cultures operate as pacers of realtime—a perspective which complicates universal accounts of realtime (Berry 2011c; Berry 2011a) and realtime research (Back, Lury, and Zimmer 2013; Lury 2012; Elmer 2013). This empirical account of realtimeness differs from a variety of current interests in realtime research. Instead of focusing on making research itself 'live', the making of what constitutes the 'live itself' is emphasized. The case study added to the previously observed distinction between 'liveness' and 'liveliness' by differentiating between 'relevance' and 'freshness' in realtime research, two device-specific modes of organizing growing amounts of information, and therefore different ways of introducing rhythm and pattern to their circulation. The configurations deployed here tap into the ordering logics of the digital media by capturing the rate of changes to their streams or result pages.

Chapter 4 introduced historical research with the Internet Archive Wayback Machine and engaged with the notion of volatility. The research affordances of the Wayback Machine are typically considered to be URL histories and evidentiary functions. In addition to repurposing the Wayback Machine's features and logics, the case study also repurposed blogging technologies and practices, informing what may be termed 'historical network analysis'. By conjuring up past states of a historical web, mainly focusing on URLs and HTML technology as a research device, the case study described the evolution of technologies and practices informing network analysis (for example, the emergence of social media buttons). In addition, this chapter laid the foundations for the volatility of methods on the web, since it turned the update culture of the Dutch blogosphere into an object of study. The historical blogosphere research focused on the software driving the Dutch blogosphere, including its defining features such as RSS fostering connections between blogs, underpinning the infrastructural changes in the blogosphere and the web more in general. The state of the web—or the periodization of the rise and decline of web technology and its cultures of use—is a research area of its own. Two larger conclusions follow from this perspective. First, the type of digital research enabled by web devices changes in sync with the update culture of the web; and second, digital methods research should be considered as flexible and adaptive.

Chapter 5 further engaged with the notion of the volatility of methods with a study into Google algorithms and their evolution, leading to new or significantly changed modes of research, or making certain research opportunities disappear altogether. The chapter focused on how repurposing as a research tactic is adaptive as opposed to principled. As discussed in relation to the notion of medium dependency, the volatility of algorithms not only has implications for the modes of research that online media enable; it also shapes the conditions of what can be known with it more in general. The volatility of capturing, formatting, and most importantly, recommendation algorithms may thus change what can be known with them. The periodization of key algorithm updates in this chapter was connected to specific modes of research contributing to both social and medium research, such as 'ranking research', 'source distance research', 'national web research', 'realtime research' and 'personalization research'. I contributed to a periodization of web research by looking at the research affordances of web devices.

Chapter 6 engaged with the device culture of the Wikipedia platform. In order to effectively repurpose a medium as a digital device, the medium's workings, central logics, and device cultures must be sufficiently understood. The case of Wikipedia is a special one; unlike most proprietary platforms and engines, Wikipedia has a transparent back-end. Its source code, policies, edit histories, and discussions are available for public scrutiny. Wikipedia's administrative apparatus and device culture, which are primed at maintaining encyclopedianess, are designed to defuse controversy, and thus aligns nicely with the pursuit of investigating the instability of knowledge in 'controversy research'. By following Wikipedia's medium-specific features and processes, the 'messiness' of Wikipedia allowed for the development of a useful tool which repurposes Wikipedia to study controversy. This chapter showed that the tactic of repurposing leads researchers to an operational definition—as opposed to an a priori definition—if they closely follow the analytical affordances of the device.

The final case study reported on in Chapter 7 introduced device cultures as a demarcation and sampling technique, by taking an approach that repurposes the significant device cultures of a national web as starting points for further research. The Iranian web was studied across a number of online media and retained the specific device cultures to compare the 'liveliness' across digital media, as measured in terms of response codes and the extent to which the webs were censored. In so doing, the device perspective contributes to censorship research by offering an approach to create a relevant sample of an otherwise unknown population of web pages.

Each of the case studies addressed key steps in the configuration of the research device, including query design, demarcation, indicators, and analysis. I have been arguing throughout this dissertation that in digital research medium research and social research overlap in an interesting way. The re-purposing debate taken up at the onset of this dissertation mainly emerged from a social research outlook, but I have been making a case for medium research operationalizing the medium as an object and analytical process in the method. The device-driven approach to digital research requires adaptive methods, but also adaptive research objectives since the digital media as vehicle for analysis are dedicated to providing novel and innovating intersections of the social and the medium for myriad purposes. The objects of study are volatile and evolve with the medium and its device cultures. Moreover, as the social reality of the ground increasingly gains traction in the study of the online and vice versa the issues troubling digital culture and social life are connected in unexpected ways, leading to the need for researchers to have their studies informed by the novel overlaps between social and medium research.

Summary

Repurposing digital methods: The research affordances of platforms and engines

This dissertation intends to contribute to current debates in digital research in two ways. First, a disciplinary contribution to the areas of computational social research and software studies is made, by developing a 'device-driven' perspective which is attentive to the operational capacities of digital media and the ways in which these media can be made productive as sources of data for digital research. Second, a contribution to digital methods is made by introducing the notion of 'research affordances' in order to provide ways to operationalize the device-driven perspective and to discuss the role of digital media as sources of data and their implications for the research process.

I examine the ways in which collections of data harvested from platforms and engines shape, and are shaped by, digital research. My line of inquiry deals with the premise that currently too little attention is paid to how digital data enclose the analytical assumptions and expectations through which it has been informed by the digital medium and its cultures of use. It is often overlooked that the digital media producing and providing access to the data sets are methodological devices, too. This in turn challenges certain assumptions and expectations brought to the data by digital researchers. Moreover, focusing on web technologies and how they operate as devices in digital methods has analytical value. They may function as an entry point into the connection between digital media, their cultures of use and the research apparatus including research questions, tools, data collection, analysis and findings.

I contribute to digital social research with a software studies informed device perspective. Much of the debate about the relations between digital method and the digi-

tal media it collects data from focuses on the term repurposing and engages with the tensions in the research process emerging from working with digital data. By demarcating this as a debate I seek to collect and contribute to questions digital methods researchers are currently examining, such as the difference between social research and medium research and the alignment between medium and method. I contribute to the repurposing debate by inquiring into the role digital platforms and engines play in digital research by means of a digital device perspective. I am drawn to the term digital device because it allows me to bring together digital media as an object and at the same time as a process in the context of digital research.

The device approach to digital media I explore, entails the cultivation of a certain medium-specific sensibility that seeks to appreciate the ways in which media objects can format and activate a diversity of practices and the way cultures of use find their conditions in the objects and forms of their media environments. The term device is thus understood as how technologies, techniques, cultures, practices and activity become operative. I thereby connect to one of the core concerns for software studies, namely the question of specific operations of software.

I also make a contribution to software studies by advancing digital methods to study software, which I refer to as 'medium research'. Methodologically, software studies are a field with a varied and rich set of approaches and methods, although new methodologies are still called for. The objects and methods of study are varied and this is perhaps the strength of the field; as software is increasingly diverse and all-pervasive, the field welcomes new methods on new objects. In addition to contributing to social research with a software studies informed device perspective, my answer to that call for new methodologies is the development of 'medium research'.

To develop the device perspective to digital research, I introduce the term 'research affordances' of digital media, focusing specifically on the analytical affordances of platforms and engines as devices in digital research. These concern the relation between objective, medium and method, and are specific to the actors and contexts of use. Research affordances of digital devices mobilize the analytical capacities of the medium and their specific device cultures. I explore the analytic qualities of the digital, and in particular the research affordances of digital platforms and engines, by bringing together a software studies perspective with a device-driven social research approach under the umbrella of digital research.

Through six case studies I empirically and conceptually explore how data outputted by a variety of devices can be used for digital research. I do so by engaging with research affordances and by addressing three contemporary issues in digital research. 1) I engage with the issue of 'device dependency' and how it can be made productive in the two perspectives which are key to digital research: the social research perspective and the medium research perspective. 2) I focus on the volatility of method in digital research, by engaging with the flexibility and indeterminacy of the various components of the methodological apparatus, as well as with the evolution of the Google algorithm and how this has afforded different modes of research over the years. 3) I focus on the notion of device cultures, introducing cultures of use, social practices and the frequency dynamics that result as key components for digital research, by an empirical exploration of the workings of Wikipedia's device culture as a key device culture in contemporary digital culture and by developing a comparative device culture analysis as a productive way to sample relevant URLs for censorship research. Let me briefly summarize the main contributions of the chapters.

Chapter 2 puts the use of digital devices as source of data into a social research perspective and focuses on the available key data collection techniques. I contribute to social research by focusing on the affordances of the studied medium and by considering how they may (or may not) be repurposed for digital social research. Such repurposing of a medium is generally brought about by connecting what and how something is captured with particular research interests, and how this allows for what I call particular modes of research. By looking at the tool building techniques for extracting data from platforms and engines—scraping and calling Application Programming Interfaces (APIs), the chapter seeks to untangle the intricate connections between collection and analysis, and between digital media and research tools. The chapter engages with the repurposing debate by empirically and conceptually exploring the difference between 'researching the medium' and 'researching the social'. The case study is used to differentiate between 'liveliness' and 'liveness' in 'realtime research', which, in brief, is the difference between measuring frequency versus measuring engagement.

Chapter 3 takes a medium research perspective, which I introduce as a form of software studies by examining what the case studies can add to the area of research. I investigate what various devices capture, format, and formalize and how they present, rank, order, and prioritize this as data. Different kinds of material are studied, ranging from patents, developer blogs and help pages, to information gathered by watch-dogs, trade press, interfaces and (default) settings, as well as particular cultures of use. I contribute to software studies by repurposing digital devices in specific configurations and by using the data outputted as material for 'medium research'. I do so by engaging empirically with the politics of realtime in various digital media. The chapter explores the making of realtime in different platforms. Based on an empirical study investigating the pace at which various digital media produce new content, the different rhythms, patterns or tempos created by the interplay of devices, users' web activities and issues are traced. What emerges are distinct forms of 'realtimeness' which are not external from, but specific to devices, organized through socio-technical arrangements and practices of use. Realtimeness thus unflattens more general accounts of the realtime web and research, and draws attention to the operations built into specific platform temporalities and the political economies of making realtime. Methodologically, this chapter contributes to software studies by advancing comparative data analysis to empirically study the effects of different realtimes created by digital media. In so doing the platforms and engines are repurposed as devices for software studies.

Chapter 4 engages with the volatility of method in digital research, by focusing on how specific configurations of digital devices, research aim and mode of analysis may affect the behavior of its components. It inquires into the procedures by which the Wayback Machine selects and reformats digital data, and looks into the extent to which it allows for other types of research besides the single site history—which the interface of the Wayback Machine privileges. The chapter proposes a methodology to repurpose the Wayback Machine so as to trace and map transitions in linking technologies and practices in a blogosphere from 1999 until 2009. By using traces of technical markers, techniques and methods for historical network analysis are introduced, and the temporal dynamics of the Dutch blogosphere are analyzed. Such an approach enables the study of the emergence and decline of blog platforms and social media platforms within the blogosphere and it enables the investigation of local blog cultures. Using historical data for research, however, also brings to light some of its limitations, such as incomplete data sets and the loss of contextual data.

Chapter 5 also engages with the volatility of methods by discussing the extent to which the update culture of key web technologies affects changing research affordances. I underline how device-driven research turns online platforms and engines—Google in this case—into a study object as well as a process structuring and changing the digital research enabled by it. By studying a variety of materials, the chapter inquires into Google's capturing and organizing logics. By regarding these logics as the conditions of possibility for what can be known with them, they are made empirically researchable. I present a temporal overview of Google's key algorithm changes and connect them to the modes of research they afford or discontinue. As such, this chapter lays the foundations for the digital methods' notion of 'search as research' and further develops the digital methods approach to software studies. This approach offers an opportunity to discuss some of the limits of digital research and their resemblance to many of the issues troubling digital social life much more in general, such as medium dependency and the black-boxing of knowledge technologies. The epistemic trouble explored by digital research is then not just a problem of social research, but rather that of the many social practices that involve the collection, management and analysis of digital social data.

Chapter 6 introduces the notion of 'device cultures' in digital research to underline how cultures of use and social practices are key components in shaping the purposes of digital devices and how that may be productively repurposed for digital research. The chapter inquires into Wikipedia's encyclopedic apparatus—a bureaucratic apparatus of policies, guidelines and essays—and connects it to the processes of knowledge production through its content management system. To maintain its 'encyclopedianess' the platform has mechanisms in place with which consensus is designed, such as the core content policies' 'neutral point of view', 'verifiability', and 'no original research'. The back-end of an article, its edit history and talk pages document the work involved in reaching consensus. In this chapter it is argued that Wikipedia's socio-technical apparatus may thus be mobilized to trace and map controversies, since its prime aim is to defuse them. In order to underpin the idea of repurposing Wikipedia for controversy research the chapter debates the quality of the digital encyclopedia's content.

Chapter 7 offers a comparative 'device culture' approach by studying the Iranian web across a variety of its popular digital media, thus conceptualizing, demarcating, and analyzing a national web. Instead of a priori defining the types of websites (as an archivist would do) to be included in a national web, the proposed methodology uses device cultures to provide (ranked) lists of URLs relevant to a particular country. These lists are subsequently studied on the basis of certain common characteristics (such as responsiveness and page age), and repurposed to study censorship in the national web. The demarcation is performed by means of different devices, which are rendered comparative; a chain of metrics is then developed in order to make claims about both the liveliness of a web and the extent to which it is censored. The chapter advanced national web research by developing ways to demarcate it and study it, and contributes to censorship research by studying its effectiveness as well as circumvention.

I conclude the dissertation by drawing together the conclusions and findings from the specific case studies. What can be considered as a good digital device is discussed in the framework of the repurposing debate. The quality of a device is not considered as an inherent property of the digital device, but rather the device becomes a good choice in relation to the research objective. My contribution to the repurposing debate slightly shifts the focus of the debate to the role of the digital device in the research process. The key concern then is how to operationalize the device for research purposes and align the various components of the research apparatus. Or, how to avoid neglecting the assumptions and norms the device brings to the research and instead to employ the media operations by configuring them in relation to the research's objective? Repurposing, operationalizing research aims through device affordances and use practices, turns digital research methods often into rather complex assemblies of the forces involved.

Nederlandse samenvatting

Herbestemming van digitale methoden: De onderzoeksmogelijkheden van platforms en zoekmachines

Deze dissertatie heeft als doelstelling op twee manieren een bijdrage te leveren aan hedendaagse debatten in digitaal onderzoek. Ten eerste wordt er een disciplinaire bijdrage geleverd aan de velden van computationeel sociaalonderzoek en software studies door een digitaal instrument-gedreven perspectief te ontwikkelen dat aandacht heeft voor de operationele capaciteiten van digitale media en de manier waarop deze media productief gemaakt kunnen worden als bronnen van data voor digitaal onderzoek. Ten tweede wordt er een bijdrage geleverd aan digitale methoden door de term 'onderzoeksmogelijkheden' te introduceren met als doel het instrument-gedreven perspectief te operationaliseren en de rol van digitale media als bronnen van data en de implicaties voor het onderzoeksprocess te bediscussiëren.

Deze dissertatie bestudeert de wijzen waarop gegevensverzamelingen afkomstig van platforms en zoekmachines digitaal onderzoek vormgeven maar er ook vorm aan ontlenen. Ik stel in het bijzonder dat er op dit ogenblik te weinig aandacht wordt besteed aan de manier waarop digitale gegevens de analytische veronderstellingen en verwachtingen bevatten die bepaald worden door het digitale medium en de gebruikscultuur ervan. Vaak wordt over het hoofd gezien dat de digitale media die gegevens leveren en er toegang toe verlenen eveneens methodologische instrumenten zijn. Op zijn beurt roept dat vragen op over bepaalde veronderstellingen en verwachtingen die digitale onderzoekers aan de gegevens toedichten. Bovendien heeft de nadruk op webtechnologieën en hoe ze werken als instrumenten in digitale methodes analytische waarde. Ze kunnen dienen als toegang tot het samenspel van digitale media, hun gebruikscultuur en het onderzoeksapparaat, met inbegrip van onderzoeksvragen, instrumenten, gegevensverzameling, analyse en resultaten.

Ik wil een bijdrage tot sociaal en cultureel onderzoek leveren door de rol van het digitale instrument in de digitale methode empirisch en conceptueel te onderzoeken. Een groot deel van het debat over de relatie tussen de digitale methode en de digitale media die gegevens leveren draait rond de term 'herbestemming' en gaat dieper in op de spanningen in het onderzoeksproces die voortvloeien uit het werk met digitale gegevens. Door dit als een debat af te lijnen probeer ik vragen te verzamelen en een bijdrage te leveren tot vragen die digitale methode onderzoekers op dit moment verkennen, zoals het verschil tussen sociaal onderzoek en mediumonderzoek en het op elkaar afstemmen van medium en methode. Ik draag bij tot het debat over herbestemming door te bestuderen welke rol digitale platforms en zoekmachine in digitaal onderzoek spelen, en dit vanuit het perspectief van het digitale instrument. De term 'digitaal instrument' spreekt mij aan omdat ik daardoor digitale media als object maar ook als proces in de context van digitaal onderzoek bijeen kan brengen.

De instrumentele benadering van digitale media die ik bestudeer, houdt in dat een zekere medium-specifieke gevoeligheid wordt gekoesterd bij de inschatting hoe objecten in bepaalde media allerlei praktijken kunnen formatteren en activeren, en hoe een gebruikscultuur haar uiting vindt in de objecten en vormen van de media waartoe ze behoort. Instrument wordt daarmee opgevat als de manier waarop technologieën, technieken, culturen, praktijken en functies werkzaam worden. Daarmee sluit ik mij aan bij één van de belangrijkste kwesties in de bestudering van software studies, namelijk de specifieke operaties van software.

Ik draag ook bij tot de bestudering van software door deze met digitale methodes te bestuderen: ik noem dit 'mediumonderzoek'. Methodologisch gezien is software studies een gebied met gevarieerde en veelomvattende benaderingen en methodes, hoewel er nog steeds vraag is naar nieuwe methodologieën. De studieobjecten en – methodes zijn velerlei, wat het onderwerp misschien wel zijn kracht geeft. Software is steeds meer wijdvertakt en alomtegenwoordig, waardoor het gebied baat heeft bij nieuwe methodes voor nieuwe objecten. Behalve mijn bijdrage tot het sociaal en cultureel onderzoek door de bestudering van software vanuit een instrumenteel perspectief, beantwoord ik de vraag naar nieuwe methodologieën met de ontwikkeling van 'mediumonderzoek'.

Om digitaal onderzoek instrumenteel te benaderen, kom ik met de term 'onderzoeksmogelijkheden' van digitale media, waarbij ik specifiek kijk naar de analytische mogelijkheden van platforms en zoekmachines, als instrument in digitaal onderzoek. Ze betreffen de relatie tussen doel, medium en methode, en zijn specifiek per actor en gebruikscontext.

Onderzoeksmogelijkheden van digitale instrumenten mobiliseren de analytische mogelijkheden van het medium en zijn specifieke instrumentele cultuur. Ik onderzoek de analytische hoedanigheden van het digitale, en in het bijzonder de onderzoeksmogelijkheden van digitale platforms en toestellen. Hiertoe verbind ik een perspectief uit de bestudering van software met een op instrumenten gebaseerde benadering van sociaal en cultureel onderzoek, onder het overkoepelend begrip digitaal onderzoek.

Aan de hand van zes case studies onderzoek ik empirisch en conceptueel hoe gegevens die worden aangeleverd door allerlei instrumenten gebruikt kunnen worden voor digitaal onderzoek. Ik bestudeer daartoe onderzoeksmogelijkheden en adresseer drie hedendaagse kwesties in digitaal onderzoek. 1) Ik bespreek de kwestie van mediumafhankelijkheid vanuit de twee invalshoeken die bepalend zijn voor digitaal onderzoek: de invalshoek van het sociaal onderzoek en die van het mediumonderzoek. 2) Ik richt mij op de vluchtigheid van de methode in digitaal onderzoek, door de flexibiliteit en de onbepaaldheid van de verschillende onderdelen van het methodologisch apparaat te bekijken, evenals de evolutie van het Google-algoritme en hoe dit over de jaren heen verschillende onderzoekswijzen heeft opgeleverd. 3) Ik leg de nadruk op het begrip platformcultuur, waarbij ik gebruikscultuur, sociale praktijken en de frequentiedynamieken die resulteren als centrale onderdelen voor digitaal onderzoek introduceer, door een empirisch onderzoek naar de werking van de platformcultuur van Wikipedia, die cruciaal is in de huidige digitale cultuur. Tevens ontwikkel ik een vergelijkende analyse van de platformcultuur als een lonende manier om relevante URLs te verzamelen voor onderzoek naar censuur. Hierna volgt een korte samenvatting van de belangrijkste bijdragen per hoofdstuk.

Hoofdstuk 2 plaatst het gebruik van digitale instrumenten als gegevensbron in het perspectief van sociaal onderzoek en licht de belangrijkste technieken van gegevensverzamelingstechnieken toe. Ik draag bij tot sociaal onderzoek door de mogelijkheden van het bestudeerde medium te belichten en te bekijken hoe ze al dan niet op digitaal sociaal onderzoek geheroriënteerd kunnen worden. Zo een herbestemming van een medium wordt over het algemeen bewerkstelligd door wat en hoe iets wordt aangeleverd te verbinden met een speciale insteek in onderzoek, en hoe daarbij speciale onderzoekswijzen, zoals ik ze noem, mogelijk zijn. Door de technieken te bekijken waarmee gegevens aan platforms en toestellen te onttrokken worden, - scraping en API's (applicatie –programmeerinterfaces) probeert het hoofdstuk de ingewikkelde verbanden tussen verzameling en analysen, en tussen digitale media en onderzoeksmiddelen bloot te leggen. Het hoofdstuk buigt zich over het herbestemmingdebat door empirisch en conceptueel het verschil na te trekken tussen onderzoek van het medium en onderzoek van het sociale. De case study wordt gebruikt om een onderscheid te maken tussen 'levendig' en 'levend' in het onderzoek van het nu, wat kort gezegd het onderscheid is tussen het meten van frequentie en het meten van betrokkenheid.

Hoofdstuk 3 vertrekt van het perspectief van mediumonderzoek, dat ik introduceer als een manier om software te bestuderen; ik bekijk wat de casestudies kunnen bijdragen aan het onderzoeksgebied. Ik onderzoek wat de verschillen instrumenten opleveren, formatteren en vormgeven en hoe ze dit als gegevens aanbieden, in volgorde zetten, ordenen en prioriteit geven. Verschillende soorten materiaal worden bestudeerd, gaande van octrooien, blogs van ontwikkelaars en helppagina's, tot informatie verzameld door waakhonden, de vakpers, interfaces en (standaard)instellingen, alsmede specifieke gebruiksculturen. Ik draag bij tot de bestudering van software door digitale instrumenten in specifieke configuraties te hergebruiken en door de geleverde gegevens als materiaal voor 'mediumonderzoek' te gebruiken. Daardoor houd ik mij empirisch bezig met de politiek van het nu in verscheidene digitale media. Het hoofdstuk bestudeert hoe het nu in de verschillende platforms gestalte krijgt. Gebaseerd op empirisch onderzoek naar de snelheid waarmee allerlei digitale media nieuwe inhoud produceren, worden de verschillende ritmes, patronen of tempo's die door het samenspel van instrumenten, activiteit van gebruikers op het web en problemen ontstaan, in kaart gebracht. Er blijken afgetekende vormen van het nu te bestaan die niet buiten de instrumenten staan maar er specifiek voor zijn, en die geordend worden door socio-technische regelingen en gebruikshandelingen. Het nu ontvouwt daarmee meer algemene beelden, alsook onderzoek, van het web in het nu, en benadrukt de werking van specifieke tijdselementen van platforms, alsmede de politieke economie van het nu. Methodologisch gezien draagt dit hoofdstuk bij tot de studie van sofware door een vergelijkende gegevensanalyse voor te stellen om zo empirisch de effecten van de verschillende vormen van het nu van digitale media te bestuderen. Zo worden platforms en zoekmachines geheroriënteerd naar instrumenten voor de bestudering van software.

Hoofdstuk 4 bespreekt de vluchtigheid van de methode in digitaal onderzoek, door te bestuderen hoe specifieke configuraties van digitale instrumenten, het onderzoeksdoel en de wijze van analyseren het gedrag van de onderdelen beïnvloeden. Het gaat nader in op de procedures waarmee de Wayback Machine digitale gegevens selecteert en herformatteert; het bekijkt ook in hoeverre ander onderzoek mogelijk is dat verdergaat dan het verleden van een enkele webpagina—het belangrijkste doel van de interface van de Wayback Machine. Het hoofdstuk stelt een methodologie voor om de Wayback Machine te heroriënteren zodat veranderende linktechnologieën in een blogosfeer van 1999 tot 2009 kunnen worden opgespoord en in kaart gebracht. Door de sporen van technische markers te gebruiken worden technieken en methodes voor historische netwerkanalyse opgevoerd en wordt de dynamiek in de tijd van de Nederlandse blogosfeer geanalyseerd. Met deze benadering kunnen de opkomst en neergang van blogplatforms en sociale mediaplatforms binnen de blogosfeer, evenals plaatselijke blogculturen bestudeerd worden. Onderzoek op basis van historische gegevens kent echter ook zijn beperkingen, zoals onvolledige gegevenssets en het verlies van contextuele gegevens.

Hoofdstuk 5 kijkt ook naar de vluchtigheid van methodes door te onderzoeken in hoeverre de updatecultuur van cruciale webtechnologieën invloed heeft op veranderende onderzoeksmogelijkheden. Ik bespreek hoe online platforms en zoekmachines—in dit geval Google—onderwerp van studie worden op basis van door instrumenten bepaald onderzoek, met daarnaast een proces dat het eruit volgende digitale onderzoek structureert en verandert. Door allerlei materiaal te besturen onderzoekt het hoofdstuk de logica waarmee Google gegevens verzamelt en ordent. Als deze logica beschouwd wordt als een mogelijke voorwaarde voor kennisverwerving, kan deze empirisch onderzocht worden. Ik geef een tijdlijn van de belangrijkste veranderingen in de Google-algoritmes en relateer ze aan het soort onderzoek dat erdoor mogelijk of onmogelijk wordt gemaakt. Daarmee legt het hoofdstuk de fundering voor wat in de digitale methodes 'zoeken als onderzoek' genoemd wordt, en ontwikkelt het de benadering van de digitale methodes in de bestudering van software. Deze benadering laat toe bepaalde beperkingen van digitaal onderzoek te bespreken, en in hoeverre ze lijken op vele van de problemen die het digitale sociale leven in het algemeen dwarszitten, zoals de afhankelijkheid van een platform en het zwarte-doosprincipe van kennistechnologieën. Het epistemologische probleem van digitaal onderzoek is daarmee niet enkel een probleem van sociaal of cultureel onderzoek, maar veeleer van de vele sociale praktijken die de verzameling, het beheer en de analyse van digitale sociale gegevens met zich meebrengen.

Hoofdstuk 6 voert het begrip 'platformcultuur' in digitaal onderzoek op, waarmee onderlijnd wordt hoe gebruiksculturen en sociale praktijken cruciale bestanddelen zijn bij het vormgeven van de oriëntatie van digitale instrumenten en hoe deze met succes kunnen worden geheroriënteerd voor digitaal onderzoek. Het hoofdstuk stelt onderzoek in naar het encyclopedisch apparaat van Wikipedia, een bureaucratisch apparaat van beleid, richtlijnen en essays—en verbindt het met de processen van kennisverwerving door de manier waarop de inhoud wordt beheerd. Om het karakter van 'encyclopedieheid' te behouden heeft het platform mechanismen ontworpen die gericht zijn op consensus, zoals de cruciale inhoudelijke beleidslijn 'neutraal standpunt', 'verifieerbaarheid' en 'geen origineel onderzoek'. De achterkant van een artikel, de bewerkingsgeschiedenis en de discussiepagina documenteren de weg naar consensus. In dit hoofdstuk wordt staande gehouden dat Wikipedia's socio-technische apparaat daarom gebruikt kan worden om controverses op te sporen en in kaart te brengen, omdat het belangrijkste doel ervan nu juist is deze glad te strijken. Ter schraging van het idee Wikipedia te heroriënteren naar controverse-onderzoek bespreekt het hoofdstuk de inhoudelijke kwaliteit van de digitale encyclopedie. De methode is erop gericht de juiste eigenschappen en indicatoren in te zetten voor het beoogde onderzoek.

Hoofdstuk 7 geeft een vergelijkende benadering van platformcultuur door het Iraanse web te bestuderen aan de hand van een aantal populaire digitale media, waardoor een nationaal web zichtbaar en geanalyseerd wordt. In plaats van eerst het soort websites te definiëren die in een nationaal web thuishoren (zoals een archivaris zou doen) gebruikt de voorgestelde methodologie platformculturen die voor een bepaald land relevante lijsten URLs (geordend) opleveren. Deze lijsten worden dan bestudeerd op basis van bepaalde gemeenschappelijke kenmerken (zoals responsiviteit en leeftijd van de pagina), en geheroriënteerd om censuur in het nationale web te bestuderen. Verschillende instrumenten die vergelijkbaar worden gemaakt zorgen voor de afbakening; daarna worden een reeks parameters ontwikkeld om te bepalen hoe levendig het web is en in hoeverre het wordt gecensureerd. De bestudering van een instrumenteel nationaal web is een specifiek voorbeeld van mediumspecifiek onderzoek met behulp van instrumenten die het nationale web bepalen.

Tot slot van deze dissertatie breng ik conclusies en resultaten van de verschillende case studies samen, met als doel een bijdrage te leveren tot digitale methodes. In deze zin betekent digitaal in methodes niet enkel dat een methode digitaal of met digitale gegevens werkt; er wordt uiteengezet hoe online platforms en zoekmachines gebruikt kunnen worden als onderzoeksinstrumenten. Wat een 'goed' digitaal instrument genoemd kan worden wordt besproken in het kader van het debat over herbestemming. De kwaliteit van het instrument is hierbij geen inherente kwaliteit, maar een productieve afstemming tussen onderzoekdoel en operaties van het medium. Mijn bijdrage tot het herbestemmingdebat verschuift de kern daarvan ietwat in de richting van de rol van het digitale instrument in het onderzoeksproces. De hamvraag is daarmee hoe het instrument voor onderzoeksdoeleinden kan worden ingezet en het de verschillende bestanddelen van het onderzoeksapparaat op elkaar kan afstemmen. De doelstelling is de analytische mogelijkheden van digitale media op een productieve manier in te zetten door ze met het oog op het onderzoek te configureren.

References

Abiteboul, S., G. Cobéna, J. Masanes, and G. Sedrati. 2002. "A First Experience in Archiving the French Web." In *Research and Advanced Technology for Digital Libraries*, edited by Maristella Agosti and Costantino Thanos, 2458:1–15. Berlin, Heidelberg: Springer Berlin Heidelberg.

Acharya, Anurag, Matt Cutts, Jeffrey Dean, Paul Haahr, Monika Henzinger, Urs Hoelzle, Steve Lawrence, Karl Pfleger, Olcan Sercinoglu, and Simon Tong. 2008. Information retrieval based on historical data. United States Patent and Trademark Office US 7,346,839 B2, filed December 31, 2003, and issued March 18, 2008.

Adamic, Lada A., and Natalie Glance. 2005. "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog." In *Proceedings of the 3rd International Workshop on Link Discovery*, 36–43. ACM.

Agamben, Giorgio. 2009. *"What Is an Apparatus?" And Other Essays*. Translated by David Kishik and Stefan Pedatella. Stanford, CA, USA: Stanford University Press.

Agre, Philip E. 1994. "Surveillance and Capture: Two Models of Privacy." *The Information Society* 10 (2): 101–27.

Ahmed, Sara. 2004. "Affective Economies." *Social Text 22* (2 79): 117–39.

Alexa. n.d. "Alexa Toolbar." *Alexa - Actionable Analytics for the Web*. http://www.alexa.com/toolbar.

Amir-Ebrahimi, Masserat. 2008. "Blogging from Qom, behind Walls and Veils." *Comparative Studies of South Asia, Africa and the Middle East 28* (2): 235–49.

Ammann, Rudolf. 2009. “Blogosphere 1998: Analysis.” Blog. *Tawawa*. November 5. http://tawawa.org/ark/2009/11/5/blogosphere-1998-analysis.html.

Anderson, Benedict. 1991. *1991: Imagined Communities: Reflections on the Origin and Spread of Nationalism.* Revised edition. London, UK and New York, NY, USA: Verso.

Anderson, Chris. 2008. “The End of Theory.” *Wired*, June 23. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

Apache Software Foundation. n.d. “Apache Storm.” *Apache Storm*. http://storm-project.net/.

Arms, W. Y. 2001. “Collecting and Preserving the Web: The Minerva Prototype.” *RLG DigiNews* 5 (2).

Arvidson, Allan, and Frans Lettenström. 1998. “The Kulturarw Project — The Swedish Royal Web Archive.” *The Electronic Library* 16 (2): 105–8.

Ashby, William Ross. 1956. *An Introduction to Cybernetics.* London, UK: Chapman & Hall.

Back, Les. 2010. “Broken Devices and New Opportunities: Re-Imagining the Tools of Qualitative Research.” NCRM Working Paper Series. London, UK: ESRC National Centre for Research Methods, Goldsmiths, University of London.

Back, Les, Celia Lury, and Robert Zimmer. 2013. “Doing Real Time Research: Opportunities and Challenges.” Working Paper. NCRM.

Back, Les, and Nirmal Puwar. 2012. “A Manifesto for Live Methods: Provocations and Capacities.” *The Sociological Review* 60 (S1): 6–17.

Badros, Gregory J, and Stephen R Lawrence. 2009. Methods and systems for personalized network searching. United States Patent and Trademark Office US 7,523,096 B2, filed December 3, 2003, and issued April 21, 2009.

Baeza-Yates, Ricardo, Carlos Castillo, and Efthimis N. Efthimiadis. 2007. “Characterization of National Web Domains.” *ACM Transactions on Internet Technology* 7 (2): 9.

Barlow, John Perry. 1996. "A Declaration of the Independence of Cyberspace." *Electronic Frontier Foundation* 8.

Barocas, Solon, Sophie Hood, and Malte Ziewitz. 2013. "Governing Algorithms: A Provocation Piece." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network.

Bar-Yossef, Ziv, Andrei Z. Broder, Ravi Kumar, and Andrew Tomkins. 2004. "Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay." In *Proceedings of the 13th International Conference on World Wide Web*, 328–37. ACM.

Baym, Nancy. 2007. "The Widgetized Self." Blog. *Online Fandom*. April 9. http://www.onlinefandom.com/archives/the-widgetized-self/.

BBC Trust. 2011. "BBC Trust Review of Impartiality and Accuracy of the BBC's Coverage of Science." *BBC Trust*. June. http://www.bbc.co.uk/bbctrust/our_work/editorial_standards/impartiality/science_impartiality.html.

———. 2014. "Trust Conclusions on the Executive Report on Science Impartiality Review Actions." *BBC Trust*. http://downloads.bbc.co.uk/bbctrust/assets/files/pdf/our_work/science_impartiality/trust_conclusions.pdf.

Bechmann, Anja. 2013. "Internet Profiling: The Economy of Data Intraoperability on Facebook and Google." *MedieKultur. Journal of Media and Communication Research* 29 (55): 19.

Beer, David. 2009. "Power through the Algorithm? Participatory Web Cultures and the Technological Unconscious." *New Media & Society* 11 (6): 985–1002.

———. 2012. "Using Social Media Data Aggregators to Do Social Research." *Sociological Research Online* 17 (3): 10.

Beer, David, and Roger Burrows. 2013. "Popular Culture, Digital Archives and the New Social Life of Data." *Theory, Culture & Society* 30 (4): 47–71.

Bekema, Vera, Liliana Bounegru, Andrea Fiore, Anne Helmond, Simon Marschall, Sabine Niederer, Bram Nijhof, Richard Rogers, and Elena Tiis. 2010. "Nationality of Issues. Rights Types." *Digital Methods Initiative Wiki*. June 30. https://wiki.digitalmethods.net/Dmi/NationalityofIssues.

Ben-David, Anat, and Hugo Huurdeman. 2014. “Web Archive Search as Research: Methodological and Theoretical Implications.” *Alexandria* 25 (1): 93–111.

Berners-Lee, Tim, James Hendler and Ora Lassila. 2001. “The Semantic Web.” Scientific American, May 2001, p. 29-37.

Berry, David M. 2011a. “Real-Time Streams and the @Cloud.” Blog. *Stunlaw*. January 23. http://stunlaw.blogspot.nl/2011/01/real-time-streams-and-cloud.html.

———. 2011b. *The Philosophy of Software: Code and Mediation in the Digital Age*. Basingstoke: Palgrave Macmillan.

———. 2011c. “Messianic Media: Notes on the Real-Time Stream.” *Stunlaw*. October 12. http://stunlaw.blogspot.com/2011/09/messianic-media-notes-on-real-time.html.

———. 2012. *Understanding Digital Humanities*. Palgrave Macmillan.

———. 2013. “The Future of European New Media Theory.” Amsterdam, NL.

Bicknell, Craig. 2000. “PointCast Coffin About to Shut.” *Wired*. March 29. http://archive.wired.com/techbiz/media/news/2000/03/35208.

Blood, Rebecca. 2004. “How Blogging Software Reshapes the Online Community.” *Communications of the ACM* 47 (12): 53–55.

Bodle, Robert. 2011. “Regimes of Sharing: Open APIs, Interoperability, and Facebook.” *Information, Communication & Society* 14 (3): 320–37.

Bollier, D. 2010. *The Promise and Peril of Big Data*. Washington, DC: Aspen Institute, Communications and Society Program.

Bolter, Jay David, and Richard Grusin. 2000. *Remediation: Understanding New Media*. Cambridge, MA, USA: MIT Press.

Borgman, Christine L. 2009. “The Digital Future Is Now: A Call to Action for the Humanities.” *Digital Humanities Quarterly* 3 (4).

Borra, Erik. 2015. *ErikBorra / Wikipedia_occupy_locations* (version 27cfa188). Web. Morph.io. Amsterdam, NL: Digital Methods Initiative. https://morph.io/ErikBorra/wikipedia_occupy_locations.

Borra, Erik, Taina Bucher, Carolin Gerlitz, Anne Helmond, and Esther Weltevrede. 2010. “One Day On the Internet Is Enough Aka Pace Online.” *Digital Methods Initiative Wiki*. October 4. https://wiki.digitalmethods.net/Dmi/OneDayOnTheInternetIsEnough.

Borra, Erik, and René König. 2013. “Googling 9/11: The Perspectives of a Search Engine on a Global Event.” presented at the Society of the Query #2, Amsterdam, NL, November 7. https://prezi.com/x5jjx3wgenoz/googling-911-the-perspectives-of-a-search-engine-on-a-global-event/.

Borra, Erik, David Laniado, Esther Weltevrede, Michele Mauri, Giovanni Magni, Tommaso Venturini, Paolo Ciuccarelli, Richard Rogers, and Andreas Kaltenbrunner. 2015. “A Platform for Visually Exploring the Development of Wikipedia Articles.” In *ICWSM ’15 - Proceedings of the 9th International AAAI Conference on Web and Social Media*.

Borra, Erik, and Bernhard Rieder. 2014. “Programmed Method: Developing a Toolset for Capturing and Analyzing Tweets.” *Aslib Journal of Information Management* 66 (3): 262–78.

Borra, Erik, and Ingmar Weber. 2012. “Political Insights: Exploring Partisanship in Web Search Queries.” *First Monday* 17 (7).

Borra, Erik, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. 2015. “Societal Controversies in Wikipedia Articles.” In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 193–96. Seoul, Republic of Korea: ACM.

Borthwick, John. 2009. “Distribution … Now.” Blog. *THINK / Musings*. May 13. http://www.borthwick.com/weblog/2009/05/13/699/.

Bowker, Geoffrey C., and Susan Leigh Star. 2000. *Sorting Things out: Classification and Its Consequences*. Inside Technology. Cambridge, MA, USA: MIT Press.

boyd, danah. 2005. “Academia and Wikipedia.” Blog. *Many to Many*. January 4. http://many.corante.com/archives/2005/01/04/academia_and_wikipedia.php.

———. 2006. “A Blogger’s Blog: Exploring the Definition of a Medium.” *Reconstruction* 6 (4). http://www.danah.org/papers/ABloggersBlog.pdf.

boyd, danah, and Kate Crawford. 2012. “Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon.” *Information, Communication & Society* 15 (5): 662–79.

Brandes, Ulrik, and Jürgen Lerner. 2008. “Visual Analysis of Controversy in User-Generated Encyclopedias.” *Information Visualization* 7 (1): 34–48.

Bright, Peter. 2014. “Yahoo Killing off Yahoo after 20 Years of Hierarchical Organization.” *Ars Technica*. September 27. http://arstechnica.com/information-technology/2014/09/yahoo-killing-off-yahoo-after-20-years-of-hierarchical-organization/.

Brin, Sergey, and Lawrence Page. 1998. “The Anatomy of a Large-Scale Hypertextual Web Search Engine.” *Computer Networks and ISDN Systems* 30 (1): 107–17.

Bross, Justus, Matthias Quasthoff, Philipp Berger, Patrick Hennig, and Christoph Meinel. 2010. “Mapping the Blogosphere with RSS-Feeds.” In *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference*, 453–60. IEEE.

Brügger, Niels. 2009. “Website History and the Website as an Object of Study.” *New Media & Society* 11 (1-2): 115–32.

———. 2010. *Web History*. New York, NY, USA: Peter Lang Publishing Inc.

Bruns, Axel. 2007. “Methodologies for Mapping the Political Blogosphere: An Exploration Using the IssueCrawler Research Tool.” *First Monday* 12 (5).

Bruns, Axel, Jean E. Burgess, Kate Crawford, and Frances Shaw. 2012. “#qldfloods And@ QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Floods.” Research Report. Brisbane QLD Australia: ARC Centre of Excellence for Creative Industries and Innovation, Queensland University of Technology.

Bruns, Axel, Lars Kirchhoff, and Thomas Nicolai. 2011. “Mapping the Australian Networked Public Sphere.” *Social Science Computer Review* 29 (3): 277–87.

Bruns, Axel, and Stefan Stieglitz. 2012. “Quantitative Approaches to Comparing Communication Patterns on Twitter.” *Journal of Technology in Human Services* 30 (3-4): 160–85.

Brussen, Bert. 2010. “Verplicht in Uw RSS-Reader: Weblogs Die Er écht Toe Doen.” Blog. *ThePostOnline*. December 28. http://www.dejaap.nl/2010/12/28/verplicht-in-uw-rss-reader-Weblogs-die-er-echt-toe-doen/.

Bucher, Taina. 2012a. “Programmed Sociality: A Software Studies Perspective on Social Networking Sites.” Ph.D. Dissertation, University of Oslo.

———. 2012b. “Want to Be on the Top? Algorithmic Power and the Threat of Invisibility on Facebook.” *New Media & Society* 14 (7): 1164–80.

———. 2013. “Objects of Intense Feeling: The Case of the Twitter APIs.” *Computational Culture: A Journal of Software Studies*, no. 3 (January).

———. 2014. “Theorizing ‘right Time’: Kairos and Algorithmic Culture.” In . Amsterdam, NL.

Buquet, Linda. 2012. “Google Venice Update – New Ranking Opportunities for Local SEO.” *Google Places Optimization Blog*. March 5. http://marketing-blog.catalystemarketing.com/google-venice-update-local-seo.html.

Buron, Florian M, Ramesh Balakrishnan, James C Norris, James R Muller, Thai Tran, and Lars E Rasmussen. 2010. Geographic coding for location search queries. United States Patent and Trademark Office US 7,747,598 B2, filed January 25, 2007, and issued June 29, 2010..

Burrows, Roger, and Nicholas Gane. 2006. “Geodemographics, Software and Class.” *Sociology* 40 (5): 793–812.

Callon, Michel, Jean-Pierre Courtial, William A. Turner, and Serge Bauin. 1983. “From Translations to Problematic Networks: An Introduction to Co-Word Analysis.” *Social Science Information* 22 (2): 191–235.

Callon, Michel, Yuval Millo, and Fabian Muniesa. 2007. *Market Devices*. Malden, MA, USA: Wiley-Blackwell.

Canada Centre for Global Security Studies, and Citizen Lab. 2011. “Casting a Wider Net: Lessons Learned in Delivering BBC Content on the Censored Internet.”

Castells, Manuel. 2000. *The Rise of the Network Society (The Information Age: Economy, Society and Culture, Volume 1)*. 2nd ed. Oxford, UK: Blackwell Publishers.

Censorship Explorer. 2012 Developed by Erik Borra, Emile Den Tex and Richard Rogers for the Digital Methods Initiative. http://tools.digitalmethods.net/beta/proxies/.

Chan, Kathy H. 2009. “I like This.” *Facebook.com/blog*. February 9. https://www.facebook.com/blog/blog.php?post=53024537130.

Chesney, Thomas. 2006. “An Empirical Examination of Wikipedia’s Credibility.” *First Monday* 11 (11).

Chun, Wendy Hui Kyong. 2008. “The Enduring Ephemeral, or the Future Is a Memory.” *Critical Inquiry* 35 (1): 148–71.

———. 2011. *Programmed Visions: Software and Memory*. Cambridge, MA, USA: MIT Press.

Clauson, Kevin A., Hyla H. Polen, Maged N. Kamel Boulos, and Joan H. Dzenowagis. 2008. “Scope, Completeness, and Accuracy of Drug Information in Wikipedia.” *Annals of Pharmacotherapy* 42 (12): 1814–21.

Contropedia Demo. 2015. Developed by Erik Borra, Esther Weltevrede, Paolo Ciuccarelli,

Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini for the Contropedia consortium. http://contropedia.net/demo.

Crawford, Kate, and Tarleton Gillespie. 2014. “What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint.” *New Media & Society*, July.

Crawford, Kate, Mary L. Gray, and Kate Miltner. 2014. “Critiquing Big Data: Politics, Ethics, Epistemology.” *International Journal of Communication* 8: 1663–72.

Cutts, Matt. 2005. “How to Write Queries.” *Matt Cutts: Gadgets, Google, and SEO*. November 9. https://www.mattcutts.com/blog/writing-google-queries/.

———. 2011. “Ten Recent Algorithm Changes.” *The Official Google Search Blog*. November 14. http://insidesearch.blogspot.com/2011/11/ten-recent-algorithm-changes.html.

———. 2012. “Search Quality Highlights: 52 Changes for April.” Blog. *The Official Google Search Blog.* May 4. http://insidesearch.blogspot.com/2012/05/search-quality-highlights-53-changes.html.

Dagdelen, Demet, Martin Feuz, Marije Rooze, Thomas Poell, and Esther Weltevrede. 2010. “Historical Controversies Now.” *Historical Controversies Now*. August 18. https://files.digitalmethods.net/var/historicalcontroversies.

Danowski, James A. 1993. “Network Analysis of Message Content.” *Progress in Communication Sciences* 12: 198–221.

Datar, Mayur, and Ashutosh Garg. 2010. Scalable user clustering based on set similarity. United States Patent and Trademark Office US 7,739,314 B2, filed August 15, 2005, and issued June 15, 2010.

Dean, Brian. 2015. “Google Ranking Factors: The Complete List.” *Backlinko*. October 1. http://backlinko.com/google-ranking-factors.

de Certeau, Michel. 1984. *The Practice of Everyday Life*. Translated by Steven Rendall. Berkeley, CA, USA: University of California Press.

Deibert, Ron, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain, eds. 2010. *Access Controlled: The Shaping of Power, Rights, and Rule in Cyberspace*. Cambridge, MA, USA: MIT Press.

Deibert, Ron, John Palfrey, Rafal Rohozinski, Jonathan Zittrain, and Janice Gross Stein, eds. 2008. *Access Denied: The Practice and Policy of Global Internet Filtering*. Cambridge, MA, USA: MIT Press.

Deibert, Ron, and Rafal Rohozinski. 2010. “Cyber Wars.” *Index on Censorship* 39 (1): 79–90.

Density Design, Digital Methods Initiative, Eurecat, and MédiaLab SciencesPo. 2015. “Contropedia.” *Contropedia.net*. http://contropedia.net/.

Digital Methods Initiative. n.d. “Digital Methods Initiative.” *Digitalmethods.net*

———. 2007. “Issue Image Analysis.” *Digital Methods Initiative Wiki.* July 15. https://wiki.digitalmethods.net/Dmi/IssueImageAnalysis.

———. 2009. “Climate Change Skeptics.” *Digital Methods Initiative Wiki*. February 13. https://wiki.digitalmethods.net/Dmi/ClimateChangeSkeptics.

Diligenti, Michelangelo, Wenxin Li, Fabio Lopiano, and Tristan G Upstill. 2012. Identification of implicitly local queries. United States Patent and Trademark Office US 8,200,694 B1, filed November 8, 2010, and issued June 12, 2012.

DMI-TCAT. 2014. Developed by Erik Borra, Bernhard Rieder and Emile Den Tex for the Digital Methods Initiative. https://github.com/digitalmethodsinitiative/dmi-tcat.

Drummond, David. 2010a. “A New Approach to China.” Blog. *Official Google Blog*. January 12. http://googleblog.blogspot.com/2010/01/new-approach-to-china.html.

———. 2010b. “An Update on China.” Blog. *Official Google Blog.* June 28. http://googleblog.blogspot.com/2010/06/update-on-china.html.

Du, Helen S., and Christian Wagner. 2006. “Weblog Success: Exploring the Role of Technology.” *International Journal of Human-Computer Studies* 64 (9): 789–98.

Eklöf, Jenny, and Astrid Mager. 2013. “Technoscientific Promotion and Biofuel Policy: How the Press and Search Engines Stage the Biofuel Controversy.” *Media, Culture & Society* 35 (4): 454–71.

Ekstrand, M. D., and J. T. Riedl. 2009. “Rv You’re Dumb: Identifying Discarded Work in Wiki Article History.” In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, 4. ACM.

Ellin, Brian. 2013. “Introducing Custom Timelines: Create Timelines of Tweets for Everyone.” *Twitter Developers*. November 12. https://web.archive.org/web/20140115135031/https://dev.twitter.com/blog/introducing-custom-timelines.

Elmer, Greg. 2013. “Live Research: Twittering an Election Debate.” *New Media & Society* 15 (1): 18–30.

Espeland, Wendy Nelson, and Michael Sauder. 2007. “Rankings and Reactivity: How Public Measures Recreate Social Worlds.” *American Journal of Sociology* 113 (1): 1–40.

Etling, Bruce, Karina Alexanyan, John Kelly, Robert Faris, John G. Palfrey, and Urs Gasser. 2010. “Public Discourse in the Russian Blogosphere: Mapping RuNet Politics and Mobilization.” SSRN Scholarly Paper ID 1698344. Rochester, NY: Social Science Research Network.

Facebook. n.d. “Facebook Platform Changelog - Documentation - Facebook for Developers.” *Facebook Developers*. https://developers.facebook.com/docs/apps/changelog.

Facebook Help. n.d. “Help Center.” *Ticker*. https://www.facebook.com/help/255898821192992/.

Festa, Paul. 2003. “Battle of the Blog.” *CNET*. August 4. https://web.archive.org/web/20090706193522/http://news.cnet.com/2009-1032-5059006.html.

Feuz, Martin, Matthew Fuller, and Felix Stalder. 2011. “Personal Web Searching in the Age of Semantic Capitalism: Diagnosing the Mechanisms of Personalisation.” *First Monday* 16 (2).

Fiorentini, Andrea. 2014. “Search Engine Optimization Matters: The Methodological and Theoretical Contribution of SEO to Web Search Studies.” M.A. thesis, Amsterdam, NL: University of Amsterdam.

Flickr. n.d. “Explore / About Interestingness.” *Flickr*. https://www.flickr.com/explore/interesting/.

Foucault, Michel. 1980. *Power/Knowledge: Selected Interviews and Other Writings, 1972-1977*. Edited by Colin Gordon. 1st American Ed edition. New York: Vintage.

Franklin, Michael, and Stan Zdonik. 1998. “‘Data in Your Face’: Push Technology in Perspective.” In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data (SIGMOD ’98)*, 27:516–19. New York, NY, USA: ACM.

Fuller, Matthew. 2003. *Behind the Blip: Essays on the Culture of Software*. Brooklyn, NY: Autonomedia.

———. 2008. *Software Studies: A Lexicon*. Leonardo. Cambridge, MA, USA: MIT Press.

Fuller, Matthew, and Andrew Goffey. 2012. *Evil Media*. Cambridge, MA, USA: MIT Press.

Galloway, Alexander R. 2004. *Protocol: How Control Exists after Decentralization.* Cambridge, MA, USA: MIT Press.

G-Atlas. 2011. Developed by Mathieu Jacomy for TIC Migrations. http://www.e-diasporas.fr/.

Gayo-Avello, Daniel. 2012. "'I Wanted to Predict Elections with Twitter and All I Got Was This Lousy Paper' -- A Balanced Survey on Election Prediction Using Twitter Data." *arXiv:1204.6441 [physics]*, April.

Geertz, Clifford. 1973. *The Interpretation of Cultures: Selected Essays*. Vol. 5019. Basic Books.

Gehl, Robert W. 2011. "The Archive and the Processor: The Internal Logic of Web 2.0." *New Media & Society* 13 (8): 1228–44.

———. 2014. *Reverse Engineering Social Media: Software, Culture, and Political Economy in New Media Capitalism*. Philadelphia, PA, USA: Temple University Press.

Gephi. 2013. Developed by Mathieu Bastian, Sebastien Heymann and Mathieu Jacomy for Médialab Sciences-Po. https://gephi.github.io/.

Gerlitz, Carolin, and Anne Helmond. 2013. "The Like Economy: Social Buttons and the Data-Intensive Web." *New Media & Society* 15 (8): 1348–65.

Gerlitz, Carolin, and Celia Lury. 2014. "Social Media and Self-Evaluating Assemblages: On Numbers, Orderings and Values." *Distinktion: Scandinavian Journal of Social Theory* 15 (2): 174–88.

Gerlitz, Carolin, and Bernhard Rieder. 2013. "Mining One Percent of Twitter: Collections, Baselines, Sampling." *M/C Journal* 16 (2).

Gerwig, Kate. 1997. "The Push Technology Rage... so What's Next?" *netWorker* 1 (2): 13–17.

Gibson, James J. 1977. "The Theory of Affordances." In *The People, Place, and Space Reader*, edited by Jen Jack Gieseking, William Mangold, Cindi Katz, Setha Low, and Susan Saegert, 56–60. Routledge.

Giles, Jim. 2005. "Internet Encyclopaedias Go Head to Head." *Nature* 438 (7070): 900–901.

Gillespie, Tarleton. 2003. “The Stories Digital Tools Tell.” In *New Media: Theories and Practices of Digitextuality*, edited by John Caldwell and Anna Everett, 304. Taylor & Francis.

———. 2010. “The Politics of ‘platforms.’” *New Media & Society* 12 (3): 347–64.

———. 2011. “Can an Algorithm Be Wrong? Twitter Trends, the Specter of Censorship, and Our Faith in the Algorithms around Us.” *Culture Digitally*.

———. 2014. “The Relevance of Algorithms.” In *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, 167–94. Cambridge, MA, USA: MIT Press.

Ginsburg, Faye D., Lila Abu-Lughod, and Brian Larkin. 2002. *Media Worlds: Anthropology on New Terrain*. Oakland, CA, USA: University of California Press.

Girard, Paul. 2011. “HyperText Corpus Initiative: How to Help Researchers Sieving the Web?” In . Velika dvorana, Slovenia. http://www.medialab.sciences-po.fr/publications/Girard-HCI.pdf.

Gitelman, Lisa. 2013. *“Raw Data” Is an Oxymoron*. Cambridge, Massachusetts ; London, England: MIT Press.

Glanz, James, and John Markoff. 2011. “U.S. Underwrites Internet Detour Around Censors Abroad.” *The New York Times*, June 12, sec. World. http://www.nytimes.com/2011/06/12/world/12internet.html.

GNIP. n.d. “GNIP.” *Gnip.com*. http://www.gnip.com.

Goffey, Andrew. 2008. “Algorithm.” In *Software Studies: A Lexicon*, edited by Matthew Fuller, 15–20. Leonardo. Cambridge, MA, USA: MIT Press.

Gold, Matthew K. 2012. *Debates in the Digital Humanities*. U of Minnesota Press.

Goldsmith, Jack L., and Tim Wu. 2006. *Who Controls the Internet?: Illusions of a Borderless World*. New York, NY, USA: Oxford University Press.

Google. n.d. “Company - Google.” *Google*. http://www.google.com/about/company/.

———. 2000. “Google Goes Global with Addition of 10 Languages.” *Google | News From Google.* May 9. http://googlepress.blogspot.nl/2000/05/google-goes-global-with-addition-of-10.html.

———. 2008. “Google Advanced Search.” *Google*. August 17. https://web.archive.org/web/20080817231534/http://www.google.com/advanced_search.

———. 2011a. “Google Transparency Report.” *Google Transparency Report*. August 25. http://www.google.com/transparencyreport/.

———. 2011b. “Company Overview.” *Google*. September 9. https://web.archive.org/web/20140920004810/http://www.google.com/about/company/.

———. 2012. *Search Quality Meeting: Spelling for Long Queries (Annotated)*. https://www.youtube.com/watch?v=JtRJXnXgE-A#t=78.

Google Scraper. 2007. Developed by Erik Borra, Koen Martens, Emile Den Tex, Richard Rogers, Sabine Niederer and Esther Weltevrede for the Digital Methods Initiative. https://tools.digitalmethods.net/beta/scrapeGoogle/.

Google Support. n.d. “Autocomplete.” *Google Support*. https://support.google.com/websearch/answer/106230?hl=en.

Govcom.org. 2007. “The IssueDramaturg.” *Issue Dramaturg*. http://issuedramaturg.issuecrawler.net/about.html.

Graham, Brad L. 1999. “Friday, September 10, 1999.” Blog. *A Personal Log of Travels and Travails on the Web. A BradLands Brand Website. Established 1998.* September 10. http://www.bradlands.com/weblog/comments/september_10_1999.

Granka, Laura A. 2010. “The Politics of Search: A Decade Retrospective.” *The Information Society* 26 (5): 364–74.

Griffith, Virgil. 2007. “WikiScanner.” *WikiScanner*. https://web.archive.org/web/20130510070028/http://wikiscanner.virgil.gr/.

Grimmelmann, James. 2009. “The Google Dilemma.” *New York Law School Law Review* 53: 939.

Gross, Ana. 2011. “The Economy of Social Data: Exploring Research Ethics as Device.” *The Sociological Review* 59 (s2): 113–29.

Haigh, Carol A. 2011. "Wikipedia as an Evidence Source for Nursing and Healthcare Students." *Nurse Education Today* 31 (2): 135–39.

Halavais, Alexander. 2008. *Search Engine Society*. Malden, MA, USA: Polity.

———. 2014. "Structure of Twitter: Social and Technical." In *Twitter and Society*, edited by Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann, 89:29–42. New York: Peter Lang.

Halavais, Alexander, and Derek Lackaff. 2008. "An Analysis of Topical Coverage of Wikipedia." *Journal of Computer-Mediated Communication* 13 (2): 429–40.

Halpin, Harry. 2012. "The Hidden History of the 'Like' Button." Amsterdam, NL: Institute of Network Cultures. http://networkcultures.org/blog/publication/unlike-us-reader-social-media-monopolies-and-their-alternatives/.

Hannak, Aniko, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. "Measuring Personalization of Web Search." In *Proceedings of the 22Nd International Conference on World Wide Web*, 527–38. WWW '13. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Hansell, Saul. 2007. "Google Keeps Tweaking Its Search Engine." *The New York Times*, June 3, sec. Business / Your Money. http://www.nytimes.com/2007/06/03/business/yourmoney/03google.html.

Häsä, Tomi. 2012. "Google Language Codes - Tomihasa." *Tomi Häsä's Pages*. June 11. https://sites.google.com/site/tomihasa/google-language-codes.

Hecht, B., and D. Gergle. 2009. "Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories." In *Proceedings of the Fourth International Conference on Communities and Technologies*, 11–20. ACM.

Helmond, Anne. 2008. "Blogging for Engines: Blogs under the Influence of Software-Engine Relations." M.A. thesis, Amsterdam, NL: University of Amsterdam.

———. 2010. "On the Evolution of Methods: Banditry and the Volatility of Methods." Blog. *Anne Helmond*. May 17. http://www.annehelmond.nl/2010/05/17/on-the-evolution-of-methods-banditry-and-the-volatility-of-methods/.

———. 2012. “Identity 2.0: Constructing Identity with Cultural Software.” Vol. 2. Amsterdam, NL.

———. 2013. “The Algorithmization of the Hyperlink.” *Computational Culture: A Journal of Software Studies* 3 (November).

———. 2015. “The Web as Platform: Data Flows in Social Media.” Ph.D. Dissertation, Amsterdam, NL: University of Amsterdam.

Heymans, Maureen, Radhika Malpani, Noam Shazeer, and Abhay Puri. 2011. Determining geographical relevance of web documents. United States Patent and Trademark Office US 8,086,690 B1, filed September 22, 2003, and issued December 27, 2011.

Highfield, Tim. 2009. “Which Way up? Reading and Drawing Maps of the Blogosphere.” *Ejournalist* 9 (1): 99–114.

Higson, Andrew. 1989. “The Concept of National Cinema.” *Screen* 30 (4): 36–47.

Hindman, Matthew. 2008. *The Myth of Digital Democracy*. Princeton, NJ, USA: Princeton University Press.

Hof, Robert D. 2009. “Betting on the Real-Time Web.” *BusinessWeek Magazine*, August 6. http://www.bloomberg.com/bw/magazine/content/09_33/b4143046834887.htm.

Holzner, Steven. 2008. *Facebook Marketing: Leverage Social Media to Grow Your Business.* Pearson Education.

Hootsuite Media. n.d. “What the Trend.” *Wthashtag.com*. http://wthashtag.com/.

Horling, Bryan, and Matthew Kulick. 2009. “Personalized Search for Everyone.” *Official Google Blog*. December 4. https://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html.

Hourihan, Meg. 2002. “What We’re Doing When We Blog.” *Web Development DevCenter.* June 13. http://archive.oreilly.com/pub/a/javascript/2002/06/13/megnut.html.

Internet Archive Wayback Machine Link Ripper. 2009. Developed by Erik Borra, Esther

Weltevrede, Anne Helmond, Michael Stevenson, Marijn De Vries Hoogerwerff and Richard Rogers for the Digital Methods Initiative. https://tools.digitalmethods.net/beta/internetArchiveWaybackMachineLinkRipper.

Introna, Lukas D, and Helen Nissenbaum. 2000. "Shaping the Web: Why the Politics of Search Engines Matters." *The Information Society* 16 (3): 169–85.

Issue Crawler. 2002. Developed by Richard Rogers, Noortje Marres, David Heath, Suzi

Wells, Marieke van Dijk, Auke Touwslager, Erik Borra, Koen Martens and Andrei Mogoutov for Govcom.org. https://issuecrawler.net.

Jansen, Bernard J., and Udo Pooch. 2001. "A Review of Web Searching Studies and a Framework for Future Research." *Journal of the American Society for Information Science and Technology* 52 (3): 235–46.

Jungherr, Andreas, Pascal Jürgens, and Harald Schoen. 2012. "Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. 'Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment.'" *Social Science Computer Review* 30 (2): 229–34.

Katz, Leo. 1953. "A New Status Index Derived from Sociometric Analysis." *Psychometrika* 18 (1): 39–43.

Keen, Andrew. 2007. *The Cult of the Amateur: How Blogs, MySpace, YouTube, and the Rest of Today's User-Generated Media Are Destroying Our Economy, Our Culture, and Our Values*. New York, NY, USA: Doubleday.

Kehoe, Colleen, Jim Pitkow, Kate Sutton, Gaurav Aggarwal, and Juan D. Rogers. 1999. "Results of GVU's Tenth World Wide Web User Survey." Atlanta, GA, USA: Graphics Visualization and Usability Center, College of Computing, Georgia Institute of Technology.

Kelly, John, and Bruce Etling. 2008. "Mapping Iran′ s Online Public: Politics and Culture in the Persian Blogosphere." 2008-01. Internet & Democracy Case Study Series. Cambridge, MA, USA: The Berkman Center for Internet & Society at Harvard University. http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Kelly&Etling_Mapping_Irans_Online_Public_2008.pdf.

Kelly, Kevin, and Gary Wolf. 1993. "PUSH!" *Wired*. http://www.wired.com/wired/archive/5.03/ff_push_pr.html.

Killroy, John. 2005. "Life Outside the Google Top Ten | Kilroy James." *Kilroy James Blog*. May 5. https://web.archive.org/web/20081003013622/http://www.kilroy-james.co.uk/2008/05/life-outside-the-google-top-ten/.

Kirschenbaum, Matthew G. 2003. "Virtuality and Vrml: Software Studies after Manovich." *The Politics of Information: The Electronic Mediation of Social Change*, 149–53.

Kitchin, Rob, and Martin Dodge. 2011. *Code/Space: Software and Everyday Life*. Software Studies. Cambridge, MA, USA: MIT Press.

Kittur, Aniket, Ed H. Chi, and Bongwon Suh. 2009. "What's in Wikipedia? Mapping Topics and Conflict Using Socially Annotated Category Structure." In *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, 1509–12. ACM.

Kittur, Aniket, and Robert E. Kraut. 2008. "Harnessing the Wisdom of Crowds in Wikipedia: Quality through Coordination." In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, 37–46. ACM.

Kittur, Aniket, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. "He Says, She Says: Conflict and Coordination in Wikipedia." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 453–62. ACM.

Koerbin, Paul. 2003. "Managing Web Archiving in Australia: A Case Study." In *Proceedings of the 3th International Web Archiving Workshop*. Norway. http://iwaw.europarchive.org/04/Koerbin.pdf.

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Sciences* 111 (24): 8788–90.

Krikorian, Raffi. 2013. "New Tweets per Second Record, and How!" *Twitter Blogs*. https://blog.twitter.com/2013/new-tweets-per-second-record-and-how.

Kumar, Ravi, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. 2004. "Structure and Evolution of Blogspace." *Communications of the ACM* 47 (12): 35–39.

Landry, Tommy. 2012. “Search Engine Market Share by Country.” *Return On Now*. June 13. http://returnonnow.com/2012/06/search-engine-market-share-country/.

Lane, Kin. n.d. “History of APIs.” *Apievangalist.com*. http://history.apievangelist.com/.

Langlois, Ganaele, and Greg Elmer. 2013. “The Research Politics of Social Media Platforms.” *Culture Machine* 14 (July): 1–17.

Langlois, Ganaele, Greg Elmer, Fenwick McKelvey, and Zachary Devereaux. 2009. “Networked Publics: The Double Articulation of Code and Politics on Facebook.” *Canadian Journal of Communication* 34 (3).

Langlois, Ganaele, Fenwick McKelvey, Greg Elmer, and K. Werbin. 2009. “Mapping Commercial Web 2.0 Worlds: Towards a New Critical Ontogenesis.” *Fibreculture* 14.

Lanier, Jaron. 2006. “Digital Maoism: The Hazards of the New Online Collectivism.” *The Edge*. May 29. http://edge.org/conversation/digital-maoism-the-hazards-of-the-new-online-collectivism.

Latour, Bruno. 1988. *Science in Action: How to Follow Scientists and Engineers Through Society*. Reprint edition. Cambridge, MA, USA: Harvard University Press.

———. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. New York, NY, USA: Oxford University Press.

———. 2007. “Learning to Navigate through Controversial Datascapes: The MACOSPOL Platform.” Science in Society.

Law, John. 2004. *After Method: Mess in Social Science Research*. London, UK: Routledge.

Law, John, and Evelyn Ruppert. 2013. “The Social Life of Methods: Devices.” *Journal of Cultural Economy* 6 (3): 229–40.

Law, John, Evelyn Ruppert, and Mike Savage. 2011. “The Double Social Life of Methods.” Working Paper 95. Milton Keynes, UK: Centre for Research on Socio-Cultural Change (CRESC) Faculty of Social Sciences, The Open University.

Lazer, David, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, et al. 2009. "Life in the Network: The Coming Age of Computational Social Science." *Science (New York, N.Y.)* 323 (5915): 721–23.

Leggetter, Phil. 2011. "Real-Time Data Delivery: HTTP Streaming Versus PubSubHubbub." Blog. *Programmable Web*. January 6. http://blog.programmableweb.com/2011/01/06/real-time-data-delivery-http-streaming-versus-pubsubhubbub/.

Leong, Susan, Teodor Mitew, Marta Celletti, and Erika Pearson. 2009. "The Question Concerning (internet) Time." *New Media & Society* 11 (8): 1267–85.

Lessig, Lawrence. 2008. *Remix: Making Art and Commerce Thrive in the Hybrid Economy*. New York, NY, USA: Penguin Press HC.

Libby, Dan. 1999. "RSS 0.91 Specification." *Rssboard.org*. October 7. http://www.rssboard.org/rss-0-9-1-netscape.

Liu, Alan. 2011. "Friending the Past: The Sense of History and Social Computing." *New Literary History* 42 (1): 1–30.

———. 2015. "N + 1: A Plea for Cross-Domain Data in the Digital Humanities." Draft. University of California, Santa Barbara, CA, USA.

Loukides, M. 2010. "What Is Data Science?" *O'Reilly Radar Report*. http://radar.oreilly.com/2010/06/what-is-data-science.html.

Lovink, Geert. 2007. "Blogging, the Nihilist Impulse." *Eurozone*. January 2. http://www.eurozine.com/articles/2007-01-02-lovink-en.html.

———. 2012. *Networks Without a Cause: A Critique of Social Media*. Cambridge, UK: Polity Press.

Lury, Celia. 2012. "Going Live: Towards an Amphibious Sociology." *The Sociological Review* 60 (S1): 184–97.

Lury, Celia, and Nina Wakeford. 2012. *Inventive Methods: The Happening of the Social.* Culture, Economy, and the Social. London, UK; New York, NY, USA: Routledge.

Mackenzie, Adrian. 1997. “The Mortality of the Virtual: Real-Time, Archive and Dead-Time in Information Networks.” *Convergence: The International Journal of Research into New Media Technologies* 3 (2): 59–71.

———. 2006. *Cutting Code: Software and Sociality*. Digital Formations 30. New York, NY, USA: Peter Lang Publishing Inc.

Magnus, P. D. 2008. “Early Response to False Claims in Wikipedia.” *First Monday* 13 (9).

Manovich, Lev. 2001. *The Language of New Media*. Cambridge, MA, USA: MIT Press.

———. 2012a. “How to Follow Software Users? (Digital Humanities, Software Studies, Big Data).” *Software Studies Initiative*. http://lab.softwarestudies.com/2012/04/new-article-lev-manovich-how-to-follow.html.

———. 2012b. “Trending: The Promises and the Challenges of Big Social Data.” In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 460–75. Minneapolis, MN, USA: University Of Minnesota Press.

———. 2012c. “Data Stream, Database, Timeline: The Forms of Social Media.” *Software Studies Initiative.* October 29. http://lab.softwarestudies.com/2012/10/data-stream-database-timeline-new.html.

———. 2013. *Software Takes Command*. New York, NY, USA and London, UK: Bloomsbury Academic.

Manovich, Lev, Jeremy Douglass, and William Huber. 2011. “Understanding Scanlation: How to Read One Million Fan-Translated Manga Pages.” *Image & Narrative* 12 (1): 206–28.

Marino, Mark C. 2006. “Critical Code Studies.” *Electronic Book Review* 4 (December).

Markham, Annette, and Nancy Baym. 2008. “Question Six: What Constitutes Quality in Qualitative Internet Research.” In *Internet Inquiry: Conversations About Method*, edited by Annette Markham and Nancy Baym, 173–98. Thousand Oaks, CA, USA: SAGE Publications.

Marres, Noortje. 2005. “No Issue, No Public: Democratic Deficits after the Displacement of Politics.” Ph.D. Dissertation, Amsterdam, NL: University of Amsterdam.

———. 2012. “The Redistribution of Methods: On Intervention in Digital Social Research, Broadly Conceived.” *The Sociological Review* 60 (S1): 139–65.

———. 2015. “Why Map Issues? On Controversy Analysis as a Digital Method.” *Science, Technology & Human Values*, March.

Marres, Noortje, and Carolin Gerlitz. 2015. “Interface Methods: Renegotiating Relations between Digital Social Research, STS and Sociology.” *Sociological Review*, February.

Marres, Noortje, and Richard Rogers. 2005. “Recipe for Tracing the Fate of Issues and Their Publics on the Web.” In *Making Things Public: Atmospheres of Democracy*, edited by Bruno Latour and Peter Weibel, 922–35. Cambridge, MA, USA: MIT Press.

Marres, Noortje, and Esther Weltevrede. 2013. “Scraping the Social? Issues in Live Social Research.” *Journal of Cultural Economy* 6 (3): 313–35.

———. 2015. “Scraping the Social? Issues in Real-Time Social Research.” In *La Médiatisation de l’évaluation/Evaluation in the Media*.

Marz, Nathan. 2011. “A Storm Is Coming: More Details and Plans for Release.” Blog. *Twitter Engineering Blog*. August 4. https://web.archive.org/web/20110810053658/http://engineering.twitter.com/2011/08/storm-is-coming-more-details-and-plans.html.

Mayer-Schönberger, Viktor. 2011. *Delete: The Virtue of Forgetting in the Digital Age*. Princeton: Princeton University Press.

Médialab SciencesPo. 2014. “Hyphe.” *GitHub*. https://github.com/medialab/hyphe.

MediaWiki Contributors. n.d. “MediaWiki.” Wiki. *MediaWiki*. https://www.mediawiki.org/w/index.php?title=MediaWiki&oldid=545966.

Meeuwsen, Frank. 2010. *Bloghelden*. Houten, NL: Bruna Uitgevers, A.W.

Metaxas, Panagiotis Takis, Eni Mustafaraj, and Daniel Gayo-Avello. 2011. “How (Not) to Predict Elections.” In *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, 165–71. IEEE.

Miller, Daniel, and Don Slater. 2001. *The Internet: An Ethnographic Approach*. Oxford, UK and New York, NY, USA: Bloomsbury Academic.

Moens, Marie-Francine. 2006. *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Vol. 21. Dordrecht, NL: Springer Netherlands.

Moretti, Franco. 2013. *Distant Reading*. London, UK; New York, NY, USA: Verso.

Moz.com. n.d. "Google Algorithm Change History." *Moz: SEO Software, Tools and Resources for Better Marketing*. https://moz.com/google-algorithm-change.

———. n.d. "Moz Rank Tracker." *Moz: SEO Software, Tools and Resources for Better Marketing*. https://moz.com/tools/rank-tracker.

Muddiman, Ashley. 2013. "Searching for the Next U.S. President: Differences in Search Engine Results for the 2008 U.S. Presidential Candidates." *Journal of Information Technology & Politics* 10 (2): 138–57.

Murdoch, Steven J., and Ross Anderson. 2008. "Tools and Technology of Internet Filtering." In *Access Denied: The Practice and Policy of Global Internet Filtering*, 52–72. Cambridge, MA, USA: MIT Press.

Murthy, Dhiraj, and Scott A. Longwell. 2013. "Twitter and Disasters: The Uses of Twitter during the 2010 Pakistan Floods." *Information, Communication & Society* 16 (6): 837–55.

Musser, John. 2005. "ProgrammableWeb: About." *Programmableweb.com*. November 24. https://web.archive.org/web/20051124200326/http://www.programmableweb.com/about.

Nanji, Ayaz. 2014. "Eye-Tracking Study: How Users View Google Search Result Pages." *MarketingProfs*. October 6. http://www.marketingprofs.com/charts/2014/26167/eye-tracking-study-how-users-view-google-search-result-pages.

National Library of the Netherlands. 2009. "KB Selectiecriteria Webarchivering."

Naumann, Sven. 2015. "Duplicate Content due to Scrapers." *Official Google Webmaster Central Blog*. October 1. http://googlewebmastercentral.blogspot.com/2008/06/duplicate-content-due-to-scrapers.html.

Nayak, Pandu. 2012. "Search Quality Highlights: 65 Changes for August and September." *The Official Google Search Blog*. October 4. http://insidesearch.blogspot.com/2012/10/search-quality-highlights-65-changes.html.

Negroponte, Nicholas. 1996. *Being Digital*. New York, NY, USA: Vintage.

Newman, Mark, Albert-László Barabási, and Duncan J. Watts. 2006. *The Structure and Dynamics of Networks*. Princeton, NJ, USA: Princeton University Press.

Niederer, Sabine, and José Van Dijck. 2010. "Wisdom of the Crowd or Technicity of Content? Wikipedia as a Sociotechnical System." *New Media & Society* 12 (8): 1368–87.

Niesyto, Johanna. 2011. "A Journey from Rough Consensus to Political Creativity: Insights from the English and German Language Wikipedias." In *Critical Point of View, a Wikipedia Reader*, edited by Geert Lovink and Nathaniel Tkacz. INC Reader 7. Amsterdam, NL: Institute of Network Cultures.

Noman, Helmi. 2008. "Tunisian Journalist Sues Government Agency for Blocking Facebook, Claims Damage for the Use of 404 Error Message instead of 403." *OpenNet Initiative*. September 12. http://opennet.net/node/950.

Norman, Donald A. 2002. *The Design of Everyday Things*. New York, NY, USA: Basic books.

OED Online. 2015. "Device, N." *OED Online*. Oxford University Press. Accessed October 21. http://www.oed.com/view/Entry/51464.

Open Net Initiative. 2009. "Internet Filtering in Iran, 2009." Research Report. Toronto: University of Toronto. https://opennet.net/sites/opennet.net/files/ONI_Iran_2009.pdf.

OpenRefine. 2013. Developed by David Huynh for Metaweb Technologies, Inc. http://openrefine.org.

O'Reilly, Tim. 2007. "What Is Web 2.0: Design Patterns and Business Models for the next Generation of Software." *Communications & Strategies*, no. 1 (March): 17–37.

Page, Lawrence. 2001. Method for Node Ranking in a Linked Database. United States Patent and Trademark Office US 6,285,999 B1, filed January 9, 1998, and issued September 4, 2001.

———. 2004. Method for scoring documents in a linked database. United States Patent and Trademark Office US 6,799,176 B1, filed July 6, 2001, and issued September 28, 2004.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. "The PageRank Citation Ranking: Bringing Order to the Web." Technical Report 1999-66. Stanford, CA, USA: Stanford Infolab.

Pasquale, Frank. 2009. "Assessing Algorithmic Authority." Blog. *Madisonian.net*. November 18. http://madisonian.net/2009/11/18/assessing-algorithmic-authority/.

Pasquale, Frank, and Oren Bracha. 2007. "Federal Search Commission? Access, Fairness and Accountability in the Law of Search." SSRN Scholarly Paper ID 1002453. Rochester, NY: Social Science Research Network.

Payne, Jason, Jake Solomon, Ravi Sankar, and Bob McGrew. 2008. "Grand Challenge Award: Interactive Visual Analytics Palantir: The Future of Analysis." In *IEEE VAST*, 201–2.

Pentland, Alex. 2014. *Social Physics: How Good Ideas Spread-The Lessons from a New Science*. Penguin.

Pinch, Trevor J., and Wiebe E. Bijker. 1984. "The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other." *Social Studies of Science* 14 (3): 399–441.

Preece, Jenny, Yvonne Rogers, Helen Sharp, David Benyon, Simon Holland, and Tom Carey. 1994. *Human-Computer Interaction*. Essex, UK, UK: Addison-Wesley Longman Ltd.

Puschmann, Cornelius, and Jean Burgess. 2014. "The Politics of Twitter Data." *Twitter and Society* 89: 43–54.

Quenqua, Douglas. 2009. "Blogs Falling in an Empty Forest." *The New York Times*, June 5, sec. Fashion & Style. http://www.nytimes.com/2009/06/07/fashion/07blogs.html.

Quick, William. 2002. *Daily Pundit*. June 11. http://web.archive.org/web/20020611194347/http://www.iw3p.com/DailyPundit/2001_12_30_dailypundit_archive.html.

Rad, Hoda Sepehri, and Denilson Barbosa. 2012. "Identifying Controversial Articles in Wikipedia: A Comparative Study." In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, 7. ACM.

Ramsey, Mike. 2012. "Understand and Rock the Google Venice Update." *Moz: SEO Software, Tools and Resources for Better Marketing*. March 12. http://moz.com/blog/understand-and-rock-the-google-venice-update.

Reagle, Joseph Michael. 2010. *Good Faith Collaboration: The Culture of Wikipedia.* Cambridge, MA, USA: MIT Press.

Recordon, David. 2009. "Tornado: Facebook's Real-Time Web Framework for Python." *Facebook Developers.* October 9. https://developers.facebook.com/blog/post/301/.

Rector, Lucy Holman. 2008. "Comparison of Wikipedia and Other Encyclopedias for Accuracy, Breadth, and Depth in Historical Articles." *Reference Services Review* 36 (1): 7–22.

Rhoads, Christopher, and Farnaz Fassihi. 2011. "Iran Vows to Unplug Internet." *Wall Street Journal,* May 28, sec. Tech. http://www.wsj.com/articles/SB10001424052748704889404576277391449002016.

Richards, Jonathan, Alex Graul, Martin Shuttleworth, Mariana Santos, Ranjit Dhaliwal, Mee-Lai Stone, and Alastair Dant. 2011. "How Twitter Tracked the News of the World Scandal." *The Guardian*. July 13. http://www.theguardian.com/media/interactive/2011/jul/13/news-of-the-world-phone-hacking-twitter.

Rieder, Bernhard. 2012. "What Is in PageRank? A Historical and Conceptual Investigation of a Recursive Status Index." *Computational Culture*, no. 2.

———. 2013. "Studying Facebook via Data Extraction: The Netvizz Application." In *Proceedings of the 5th Annual ACM Web Science Conference*, 346–55. ACM.

———. 2015. "The End of Netvizz (?)." *The Politics of Systems*. Accessed October 1. http://thepoliticsofsystems.net/2015/01/the-end-of-netvizz/.

Rieder, Bernhard, and Theo Röhle. 2012. "Digital Methods: Five Challenges." In *Understanding Digital Humanities*, 67–84. London, UK: Palgrave Macmillan.

Roberts, Hal, Ethan Zuckerman, and John Palfrey. 2011. “2011 Circumvention Tool Evaluation.” SSRN Scholarly Paper ID 1940455. Rochester, NY: Social Science Research Network.

Roblimo. 2004. “Wikipedia Founder Jimmy Wales Responds.” *Slashdot*. July 28. http://beta.slashdot.org/story/04/07/28/1351230/wikipedia-founder-jimmy-wales-responds.

Rogers, Richard. 2002. “Towards a Live Social Science on the Web.” *EASST Review* 21 (3/4): 2–4.

———. 2004. “Why Map? The Techno-Epistemological Outlook.” Essay. Media Design Research. Rotterdam, NL: Piet Zwart Institute, Willem de Kooning Academie.

———. 2005. “Old and New Media: Competition and Political Space.” *Theory & Event* 8 (2).

———. 2009. *The End of the Virtual: Digital Methods*. Vol. 339. Amsterdam, NL: Amsterdam University Press.

———. 2012. “Mapping and the Politics of Web Space.” *Theory, Culture & Society* 29 (4-5): 193–219.

———. 2013a. “Debanalizing Twitter: The Transformation of an Object of Study.” In *Proceedings of the 5th Annual ACM Web Science Conference*, 356–65. ACM.

———. 2013b. *Digital Methods*. Cambridge, MA, USA: MIT Press.

Rogers, Richard, Fieke Jansen, Michael Stevenson, Esther Weltevrede, and A. Finlay. 2009. “Mapping Democracy.” Global Informaton Society Watch 2009. Association for Progressive Communications and Hivos.

Rogers, Richard, Natalia Sánchez-Querubín, and Aleksandra Kil. 2015. *Issue Mapping for an Ageing Europe*. Amsterdam University Press.

Rogers, Richard, Esther Weltevrede, Sabine Niederer, and Erik Borra. 2011. “National Web Studies: Mapping Iran Online.” *Digitalmethods.net*. http://mappingiranonline.digitalmethods.net.

———. 2013. “National Web Studies: The Case of Iran Online.” In *A Companion to New Media Dynamics*, 142–66. Oxford, UK: Wiley-Blackwell.

Rosenberg, Scott. 2009. *Say Everything: How Blogging Began, What It's Becoming, and Why It Matters*. Broadway Books.

Ruppert, Evelyn. 2007. "Producing Population." Working Paper. Milton Keynes, UK: Centre for Research on Socio-Cultural Change (CRESC) Faculty of Social Sciences, The Open University.

Ruppert, Evelyn, John Law, and Mike Savage. 2013. "Reassembling Social Science Methods: The Challenge of Digital Devices." *Theory, Culture & Society* 30 (4): 22–46.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms." In . Seattle, WA, USA.

Sanger, Lawrence M. 2001. "Wikipedia Policy." *Webcitation.org*. January 11. http://www.webcitation.org/5U0Rx8xAc.

———. 2009. "The Fate of Expertise after Wikipedia." *Episteme* 6 (1): 52–73.

Sarawagi, Sunita. 2007. "Information Extraction." *Foundations and Trends® in Databases* 1 (3): 261–377.

Savage, Mike. 2010. *Identities and Social Change in Britain since 1940: The Politics of Method*. Oxford, UK; New York, NY, USA: Oxford University Press.

Savage, Mike, and Roger Burrows. 2007. "The Coming Crisis of Empirical Sociology." *Sociology* 41 (5): 885–99.

Savage, Mike, Evelyn Ruppert, and John Law. 2010. "Digital Devices: Nine Theses." Working Paper 86. CRESC Working Paper Series. Manchesterm UK: CRESC, The Open University of Manchester.

Schaap, Frank. 2004. "Links, Lives, Logs: Presentation in the Dutch Blogosphere." Article. Into the Blogosphere Articles. Minneapolis, MN, USA: University of Minnesota.

Schmidt, Eric. 2009. "Prosperity or Peril? The Next Phase of Globalization." In *Princeton Colloquium on Public and International Affairs*. Vol. 18. Princeton, NJ, USA.

———. 2011. “Testimony of Eric Schmidt, Executive Chairman, Google Inc. Before the Senate Committee on the Judiciary Subcommittee on Antitrust, Competition Policy, and Consumer Rights.” Washington, DC, USA. http://searchengineland.com/figz/wp-content/seloads/2011/09/Eric-Schmidt-Testimony.pdf.

Schmidt, Jan. 2007. “Blogging Practices: An Analytical Framework.” *Journal of Computer-Mediated Communication* 12 (4): 1409–27.

Schwartz, Barry. 2014. “Google Blog Search Now Within Google News Search.” *Search Engine Land*. August 29. http://searchengineland.com/google-blog-search-now-within-google-news-search-202202.

ScraperWiki. n.d. “Classic Scraper Wiki.” *ScraperWiki.* https://classic.scraperwiki.com/.

Searls, Doc. 2005. “Linux for Suits: The World Live Web.” *Linux Journal,* October 31.

Shah, Chirag, and Charles File. 2011. “InfoExtractor – A Tool for Social Media Data Mining.” *JITP 2011: The Future of Computational Social Science* 7 (January): 24.

Shaheed, Ahmed. 2014. “Situation of Human Rights in the Islamic Republic of Iran.” Situations and reports of special rapporteurs and representatives A/69/356. New York, NY, USA: United Nations, General Assembly.

Shaw, Aaron, and Yochai Benkler. 2012. “A Tale of Two Blogospheres: Discursive Practices on the Left and Right.” *American Behavioral Scientist* 56 (4): 459–87.

Sherman, Chris. 2005. “Google Personalized Search Leaves Google Labs.” Blog. *Search Engine Watch*. November 9. http://searchenginewatch.com/sew/news/2067833/google-personalized-search-leaves-google-labs.

SIDN. 2007. “SIDN Kondigt Uitfasering Persoonsdomeinnamen Aan.” *SIDN.nl*. https://www.sidn.nl/nieuws/nieuwsbericht/article/sidn-kondigt-uitfasering-persoonsdomeinnamen-aan/.

———. 2010. “Jaarverslag 2010: Het Jaar Dat Internet Het Nieuws Beheerste.” https://www.sidn.nl/fileadmin/docs/PDF-files_NL/SIDN_Jaarverslag_2010.pdf.

Singhal, Amit. 2009. “Relevance Meets the Real-Time Web.” Blog. *Official Google Blog*. December 7. http://googleblog.blogspot.com/2009/12/relevance-meets-real-time-web.html.

———. 2011. “Giving You Fresher, More Recent Search Results.” *Official Google Blog*. November 3. https://googleblog.blogspot.com/2011/11/giving-you-fresher-more-recent-search.html.

———. 2012. “Search Quality Highlights: 40 Changes for February.” *Inside Search. The Official Google Search Blog*. February 27. http://insidesearch.blogspot.com/2012/02/search-quality-highlights-40-changes.html.

Slawski, Bill. 2014. “Search Queries.” *SEO by the Sea*. October 27. http://www.seobythesea.com/category/searchers-queries/.

Slee, Tom. 2011. “Internet-Centrism 3 (of 3): Tweeting the Revolution (and Conflict of Interest).” *Whimsley*. September 22. http://whimsley.typepad.com/whimsley/2011/09/earlier-today-i-thought-i-was-doomed-to-fail-that-part-3-of-this-prematurely-announced-trilogy-was-just-not-going-to-get-wr.html.

Sobek, Markus. 2002. “Google Dance - The Index Update of the Google Search Engine.” *Efactory.de.* http://dance.efactory.de/.

Solomon, Susan, Dahe Qin, Martin Manning, Zhenlin Chen, Melinda Marquis, Kristen Averyt, Melinda M.B. Tignor, and Henrey LeRoy Miller, eds. 2007. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge, UK and New York, NY, USA: Cambridge University Press.

Source Code Search. 2012. Developed by Erik Borra, Esther Weltevrede and Anne Helmond for the Digital Methods Initiative. http://tools.digitalmethods.net/beta/sourceCodeSearch.

Stalder, Felix. 2012. “Between Democracy and Spectacle: The Front-End and Back-End of the Social Web.” In *The Social Media Reader,* edited by Michael Mandiberg, 242–56. New York, NY, USA: NYU Press.

Stalder, Felix, and Christine Mayer. 2009. “The Second Index. Search Engines, Personalization and Surveillance.” In *Deep Search: The Politics of Search beyond Google*, edited by Konrad Becker and Felix Stalder, 98–115. Innsbruck, AT; Piscataway, NJ, USA: Studien Verlag.

Stanfill, Mel. 2015. “The Interface as Discourse: The Production of Norms through Web Design.” *New Media & Society* 17 (7): 1059–74.

Stankus, Tony, and Sarah Spiegel. 2010. "Wikipedia, Scholarpedia, and References to Journals in the Brain and Behavioral Sciences: A Comparison of Cited Sources and Recommended Readings in Matching Free Online Encyclopedia Entries." *Science & Technology Libraries* 29 (3): 258–65.

Steinbrenner, Karin. 2001. "Unlocking ERPs with Portals." *Educause Quarterly*, no. 3 (January): 55–57.

Stevenson, Michael. 2010. "The Archived Blogosphere: Exploring Web Historical Methods Using the Internet Archive." Amsterdam, NL.

———. 2013. "The Web as Exception: The Rise of New Media Publishing Cultures." Ph.D. Dissertation, Amsterdam, NL: University of Amsterdam.

Suh, Bongwon, Ed H. Chi, Bryan A. Pendleton, and Aniket Kittur. 2007. "Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations." In *VAST 2007. IEEE Symposium on Visual Analytics Science and Technology, 2007.*, 163–70. IEEE.

Sullivan, Danny. 2007a. "Google Kills Bush's Miserable Failure Search & Other Google Bombs." *Search Engine Land*. January 25. http://searchengineland.com/google-kills-bushs-miserable-failure-search-other-google-bombs-10363.

———. 2007b. "Google Search History Expands, Becomes Web History." *Search Engine Land*. April 19. http://searchengineland.com/google-search-history-expands-becomes-web-history-11016.

———. 2010. "Dear Bing, We Have 10,000 Ranking Signals To Your 1,000. Love, Google." *Search Engine Land*. November 11. http://searchengineland.com/bing-10000-ranking-signals-google-55473.

———. 2011. "As Deal With Twitter Expires, Google Realtime Search Goes Offline." *Search Engine Land*. July 4. http://searchengineland.com/as-deal-with-twitter-expires-google-realtime-search-goes-offline-84175.

Sumi, R., T. Yasseri, A. Rung, A. Kornai, and J. Kertész. 2011. "Characterization and Prediction of Wikipedia Edit Wars."

Svensson, Patrik, and David Theo Goldberg. 2015. *Between Humanities and the Digital*. MIT Press.

Technorati. 2011. “Blog Quality Guidelines.” Blog. *Technorati.* https://web.archive.org/web/20110824003955/http://technorati.com/blog-quality-guidelines-faq.

Thelwall, Mike, Rudy Prabowo, and Ruth Fairclough. 2006. “Are Raw RSS Feeds Suitable for Broad Issue Scanning? A Science Concern Case Study.” *Journal of the American Society for Information Science and Technology* 57 (12): 1644–54.

TLD Counts. 2013. Developed by Erik Borra and Emile Den Tex for the Digital Methods Initiative. https://tools.digitalmethods.net/beta/tldCounts/.

Tracker Tracker. 2011. Developed by Koen Martens, Emile Den Tex, Erik Borra, Esther Weltevrede, Anne Helmond and Carolin Gerlitz for the Digital Methods Initiative. https://tools.digitalmethods.net/beta/trackerTracker/.

Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.” *ICWSM* 10: 178–85.

Twitter. n.d. “Twitter About.” *Twitter About*. https://about.twitter.com/.

———. 2010. “Twitter 2010: Year in Review.” *Twitter Year In Review*. http://yearinreview.twitter.com/2010/trends/.

Twitter Developers. n.d. “The Streaming APIs.” *Twitter Developers*. https://dev.twitter.com/streaming/overview.

———. n.d. “Twitter Developers.” *Twitter Developers*. https://dev.twitter.com/.

———. 2015. “Things Every Developer Should Know.” *Twitter Developers*. Accessed October 1. https://dev.twitter.com/overview/general/things-every-developer-should-know.

Twitter Support. n.d. “Getting Started with Twitter.” *Twitter Support*. https://support.twitter.com/articles/215585.

Uprichard, Emma. 2012. “Being Stuck in (live) Time: The Sticky Sociological Imagination.” *The Sociological Review* 60 (S1): 124–38.

———. 2013. “Sampling: Bridging Probability and Non-Probability Designs.” *International Journal of Social Research Methodology* 16 (1): 1–11.

Uprichard, Emma, Roger Burrows, and David Byrne. 2008. "SPSS as an 'inscription Device': From Causality to Description?" *The Sociological Review* 56 (4): 606–22.

Van Couvering, Elizabeth. 2007. "Is Relevance Relevant? Market, Science, and War: Discourses of Search Engine Quality." *Journal of Computer-Mediated Communication* 12 (3): 866–87.

Van den Berg, Walter. 2009. "Vandenb.com - Pagina 12 van 18 - Weblog van Walter van Den Berg." *Vandenb.com*. November 2. http://vandenb.com/page/12/.

Van der Velden, Lonneke. 2014. "The Third Party Diary: Tracking the Trackers on Dutch Governmental Websites." *NECSUS. European Journal of Media Studies* 3 (1): 195–217.

Van Dijck, José. 2011. "Tracing Twitter: The Rise of a Microblogging Platform." *International Journal of Media & Cultural Politics* 7 (3): 333–48.

———. 2013. *The Culture of Connectivity: A Critical History of Social Media.* Oxford, UK; New York, NY, USA: Oxford University Press.

Van Dijck, José, and Thomas Poell. 2013. "Understanding Social Media Logic." SSRN Scholarly Paper ID 2309065. Rochester, NY: Social Science Research Network.

Van Ess, Henk. 2005. "Google Confirms: Eval.google.com Exists - Henk van Ess's Search Bistro." Blog. *Henk van Ess' Search Bistro*. June 6. https://web.archive.org/web/20070206060436/http://www.searchbistro.com/index.php?/archives/30-Google-Confirms-Eval.google.com-Exists.html.

Van Hoboken, Joris. 2012. "Search Engine Freedom: On the Implications of the Right to Freedom of Expression for the Legal Governance of Web Search Engines." Ph.D. Dissertation, Amsterdam, NL: University of Amsterdam.

Venturini, Tommaso. 2010. "Diving in Magma: How to Explore Controversies with Actor-Network Theory." *Public Understanding of Science* 19 (3): 258–73.

Venturini, Tommaso, and Bruno Latour. 2010. "The Social Fabric: Digital Traces and Quali-Quantitative Methods." In *Proceedings of Future En Seine 2009*, 87–101. Cap Digital.

Viégas, Fernanda, Martin Wattenberg, Jesse Kriss, and Frank Van Ham. 2007. "Talk Before You Type: Coordination in Wikipedia." In *40th Annual Hawaii International Conference on System Sciences, 2007. HICSS 2007*, 78–78. IEEE.

Vieweg, Sarah, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 2010. "Microblogging during Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 1079–88. ACM.

Villeneuve, Nart. 2006. "Testing through Proxies in China." Blog. *Malware Explorer: Nart Villeneuve*. April 10. https://www.nartv.org/2006/04/10/testing-through-proxies-in-china/.

Virilio, Paul. 1995. *Art Of The Motor*. Minnesota, MP, USA: University of Minnesota Press.

Von Hilgers, Philipp. 2011. "The History of the Black Box: The Clash of a Thing and Its Concept." *Cultural Politics* 7 (1): 41–58.

Vrints, Eugeen. 2002. "Eug's Weblog." October 7. http://web.archive.org/web/20021009224346/http://users.pandora.be/vrints/.

Waters, Neil. 2007. "Why You Can't Cite Wikipedia in My Class." *Communications of the ACM* 50 (9): 15–17.

Watters, Audrey. 2011. "Scraping, Cleaning, and Selling Big Data." *O'Reilly Radar*. May 11. http://radar.oreilly.com/2011/05/data-scraping-infochimps.html.

Wayback Network Per Year. 2010. Developed by Erik Borra, Esther Weltevrede and Anne Helmond for the Digital Methods Initiative. https://tools.digitalmethods.net/beta/waybackNetworkPerYear.

Webber, Richard. 2009. "Response to 'The Coming Crisis of Empirical Sociology': An Outline of the Research Potential of Administrative and Transactional Data." *Sociology* 43 (1): 169–78.

Weber, Ingmar, and Carlos Castillo. 2010. "The Demographics of Web Search." In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 523–30. ACM.

Weltevrede, Esther. 2009. “Thinking Nationally with the Web. A Medium-Specific Approach to the National Turn in Web Archiving.” M.A. thesis, Amsterdam, NL: University of Amsterdam.

———. 2010. “Studying Society, Not Google?” Amsterdam, NL.

———. 2015a. *The Research Browser*. New Media Research Methods. Amsterdam, NL: Digital Method Initiative. https://www.youtube.com/watch?v=bj65Xr9GkJM.

———. 2015b. “Google Algorithm Changes.” Wiki. *Digital Methods Initiative Wiki*. August 11. https://wiki.digitalmethods.net/pub/Dmi/RightsTypes/Updates_timeline_.pdf.

Weltevrede, Esther, and Erik Borra. 2015. “Controversy in the Back-End of Neutral Point of View: The Research Affordances of Wikipedia for Studying Societal Issues.” Unpublished ms.

Weltevrede, Esther, and Anne Helmond. 2011. “User Pages - Main Graph of NEW 2000.” Project page. *Dutch Blogosphere*. https://dutchblogosphere.digitalmethods.net/gatlas/index.php?focus=nodeattribute&graph=18&nodeattribute=6&sec tion=18.

———. 2012a. “Where Do Bloggers Blog? Platform Transitions in the Historical Dutch Blogosphere.” http://dutchblogosphere.digitalmethods.net.

———. 2012b. “Where Do Bloggers Blog? Platform Transitions within the Historical Dutch Blogosphere.” *First Monday* 17 (2).

Weltevrede, Esther, Anne Helmond, and Carolin Gerlitz. 2014. “The Politics of Real-Time: A Device Perspective on Social Media Platforms and Search Engines.” *Theory, Culture & Society* 31 (6): 125–50.

Whetstone, Rachel. 2007. “Free Expression and Controversial Content on the Web.” Blog. *Official Google Blog*. November 14. http://googleblog.blogspot.com/2007/11/free-expression-and-controversial.html.

———. 2010. “Controversial Content and Free Expression on the Web: A Refresher.” Blog. *Official Google Blog*. April 19. http://googleblog.blogspot.com/2010/04/controversial-content-and-free.html.

Widrich, Leo. 2012. “How To Tweet At The Best Times For Your Followers - Tweriod & Buffer Team up.” *The Buffer Blog*. April 11. https://blog.bufferapp.com/how-to-tweet-at-the-best-times-for-your-followers-tweriod-buffer-team-up.

Wiener, Norbert. 1961. *Cybernetics; Or, Control and Communication in the Animal and the Machine.* Vol. 25. The @MIT Paperback Series. Cambridge, MA, USA: MIT Press.

Wiggins, Richard W. 2001. “The Effects of September 11 on the Leading Search Engine.” *First Monday* 6 (10). http://journals.uic.edu/ojs/index.php/fm/article/view/890.

Wikipedia contributors. 2007a. “Global Warming.” *Wikipedia, the Free Encyclopedia.* January 9. https://en.wikipedia.org/w/index.php?title=Global_warming&oldid=99492775.

———. 2007b. “Global Warming.” *Wikipedia, the Free Encyclopedia.* February 4. https://en.wikipedia.org/w/index.php?title=Global_warming&oldid=105516662.

———. 2007c. “Global Warming.” *Wikipedia, the Free Encyclopedia.* March 25. https://en.wikipedia.org/w/index.php?title=Global_warming&oldid=117768478.

———. 2007d. “Global Warming.” *Wikipedia, the Free Encyclopedia.* December 24. https://en.wikipedia.org/w/index.php?title=Global_warming&oldid=179874419.

———. 2007e. “List of Scientists Opposing the Mainstream Scientific Assessment of Global Warming.” *Wikipedia, the Free Encyclopedia.* December 29. https://en.wikipedia.org/w/index.php?title=List_of_scientists_opposing_the_mainstream_scientific_assessment_of_global_warming&oldid=180784320.

———. 2010a. “Talk:Global warming/Archive 24#POV_in_the_intro.” *Wikipedia, the Free Encyclopedia.* September 19. https://en.wikipedia.org/w/index.php?title=Talk:Global_warming/Archive_24&oldid=385796965#POV_in_the_intro.

———. 2010b. “Talk:Global warming/Archive 55#Text_discussion.2C_2nd_section.” *Wikipedia, the Free Encyclopedia.* September 19. https://en.wikipedia.org/w/index.php?title=Talk:Global_warming/Archive_55&oldid=385799721#Text_discussion.2C_2nd_section.

———. 2010c. "Talk:List of Scientists Opposing the Mainstream Scientific Assessment of Global warming/Archive 11#title." *Wikipedia, the Free Encyclopedia.* September 19. https://en.wikipedia.org/w/index.php?title=Talk:List_of_scientists_opposing_the_mainstream_scientific_assessment_of_global_warming/Archive_11&oldid=385814601#title.

———. 2010d. "Talk:List of Scientists Opposing the Mainstream Scientific Assessment of Global warming/Archive 4." *Wikipedia, the Free Encyclopedia.* October 12. https://en.wikipedia.org/w/index.php?title=Talk:List_of_scientists_opposing_the_mainstream_scientific_assessment_of_global_warming/Archive_4&oldid=390253739.

———. 2010e. "Talk:List of Scientists Opposing the Mainstream Scientific Assessment of Global warming/Archive 4#AEB2." *Wikipedia, the Free Encyclopedia.* October 12. https://en.wikipedia.org/w/index.php?title=Talk:List_of_scientists_opposing_the_mainstream_scientific_assessment_of_global_warming/Archive_4&oldid=390253739#AEB2.

———. 2012a. "Country Code Top-Level Domain." *Wikipedia, The Free Encyclopedia.* January 26. https://en.wikipedia.org/w/index.php?title=Country_code_top-level_domain&oldid=473401638.

———. 2012b. "Wikipedia:Policies and Guidelines." *Wikipedia, the Free Encyclopedia.* December 19. https://en.wikipedia.org/w/index.php?title=Wikipedia:Policies_and_guidelines&oldid=528865694.

———. 2013a. "Wikipedia:Edit Warring." *Wikipedia, the Free Encyclopedia.* January 8. https://en.wikipedia.org/w/index.php?title=Wikipedia:Edit_warring&oldid=531927086.

———. 2013b. "Help:Reverting." *Wikipedia, the Free Encyclopedia.* January 9. https://en.wikipedia.org/w/index.php?title=Help:Reverting&oldid=532118767.

———. 2013c. "Wikipedia:User Access Levels." *Wikipedia, the Free Encyclopedia.* January 13. https://en.wikipedia.org/w/index.php?title=Wikipedia:User_access_levels&oldid=532841128.

———. 2013d. "Wikipedia:Neutral Point of View." *Wikipedia, the Free Encyclopedia.* January 14. https://en.wikipedia.org/w/index.php?title=Wikipedia:Neutral_point_of_view&oldid=533113104.

———. 2014a. “Wikipedia:Automatic Edit Summaries.” *Wikipedia, the Free Encyclopedia.* May 26. https://en.wikipedia.org/w/index.php?title=Wikipedia:Automatic_edit_summaries&oldid=610242782.

———. 2014b. “Wikipedia:Purpose.” *Wikipedia, the Free Encyclopedia.* July 2. https://en.wikipedia.org/w/index.php?title=Wikipedia:Purpose&oldid=615280437.

———. 2014c. “Wikipedia:Neutral Point of View.” *Wikipedia, the Free Encyclopedia.* August 5. https://en.wikipedia.org/w/index.php?title=Wikipedia:Neutral_point_of_view&oldid=619907416.

———. 2014d. “Wikipedia:No Original Research.” *Wikipedia, the Free Encyclopedia.* August 14. https://en.wikipedia.org/w/index.php?title=Wikipedia:No_original_research&oldid=621190554.

———. 2014e. “Wikipedia:Manual of Style/Words to Watch.” *Wikipedia, the Free Encyclopedia.* August 16. https://en.wikipedia.org/w/index.php?title=Wikipedia:Manual_of_Style/Words_to_watch&oldid=621505551.

———. 2014f. “Wikipedia:List of Controversial Issues.” *Wikipedia, the Free Encyclopedia.* August 19. https://en.wikipedia.org/w/index.php?title=Wikipedia:List_of_controversial_issues&oldid=621866745.

———. 2014g. “Wikipedia:Consensus.” *Wikipedia, the Free Encyclopedia.* August 30. https://en.wikipedia.org/w/index.php?title=Wikipedia:Consensus&oldid=623509425.

———. 2014h. “Wikipedia:Verifiability.” *Wikipedia, the Free Encyclopedia.* September 9. https://en.wikipedia.org/w/index.php?title=Wikipedia:Verifiability&oldid=624874142.

Winer, Dave. 1997. “Scripting News in XML.” *Scripting News.* December 15. http://scripting.com/davenet/1997/12/15/scriptingNewsInXML.html.

———. 2000. “What to Do about RSS?” *Scripting News.* September 2. http://scripting.com/davenet/2000/09/02/whatToDoAboutRss.html.

Wright, Joss, Tulio De Souza, and Ian Brown. 2011. “Fine-Grained Censorship Mapping: Information Sources, Legality and Ethics.” San Francisco, CA, USA: USENIX Association.

Yamaoka, So, Lev Manovich, Jeremy Douglass, and Falko Kuester. 2011. "Cultural Analytics in Large-Scale Visualization Environments." *Computer*, no. 99 (December).

Yasseri, Taha, Robert Sumi, András Rung, Andras Kornai, and János Kertész. 2012. "Dynamics of Conflicts in Wikipedia." *PLoS ONE* 7 (6): e38869.

Yegge, Steve. 2012. "Stevey's Google Platforms Rant." *Plus.google.com*. December 10. https://plus.google.com/+RipRowan/posts/eVeouesvaVX.

Yehoshua, Tamar, and Bobby Nath. 2015. "Google's Look, Evolved." Blog. *Official Google Blog.* September 1. https://googleblog.blogspot.com/2015/09/google-update.html.

Zittrain, Jonathan. 2008. *The Future of the Internet and How to Stop It*. New Haven, CT, USA: Yale University Press.

Zittrain, Jonathan, and John Palfrey. 2008. "Internet Filtering: The Politics and Mechanisms of Control." *Access Denied: The Practice and Policy of Global Internet Filtering* 41.