Scraping the Social? Issues in real-time social research

Draft Paper (March 2012), submitted for publication to the Journal of Cultural Economy

Noortje Marres, Goldsmiths, University of London
Esther Weltevrede, University of Amsterdam

Abstract

What makes scraping methodologically interesting for social and cultural research? This paper seeks to contribute to debates about digital social research by exploring how a 'medium-specific' technique for online data capture may be rendered analytically productive for social research. As a device that is currently being imported into social research, scraping has the capacity to re-structure social research, and this in at least two ways. Firstly, as a technique that is not native to social research, scraping risks to introduce 'alien' methodological assumptions into social research (such as an pre-occupation with freshness). Secondly, to scrape is to risk importing into our inquiry categories that are prevalent in the social practices enabled by the media: scraping makes available *already formatted* data for social research. Scraped data, and online social data more generally, tend to come with 'external' analytics already built-in. This circumstance is often approached as a 'problem' with online data capture, but we propose it may be turned into virtue, insofar as data formats that have currency in the areas under scrutiny may serve as a source of social data themselves. Scraping, we propose, makes it possible to render traffic between the object and process of social research analytically productive. It enables a form of 'real-time' social research, in which the formats and life cycles of online data may lend structure to the analytic objects and findings of social research. By way of a conclusion, we demonstrate this point in an exercise of online issue profiling, and more particularly, by relying on Twitter to profile the issue of 'austerity'. Here we distinguish between two forms of real-time research, those dedicated to monitoring *live* content (which terms are current?) and those concerned with analysing the *liveliness* of issues (which topics are happening?).

Keywords

Real-time research, digital social research, science and technologies studies, digital methods, automated information extraction, meta-data, information formats

1. Introduction

Scraping, to state this quite formally, is a currently prominent technique for the automated collection of online data. It is one of the more distinctive techniques associated with present forms of digital social research, those that have been grouped under the rubric of the 'second computational turn' in the social and cultural sciences, which is marked by the rise of the Internet, and is to be distinguished from a 'first' computational turn, which is associated with the rise of computational statistics in the late 1960s, and which then inspired the further development of quantitative methods in social science (Rogers forthcoming; Burrows, Uprichard et al 2008). As a technique for collecting data on the Web, scraping is widely seen to offer opportunities for developing new forms of data collection, analysis, and visualisation, and these opportunities have in recent years been examined and advertised in various programmatic pronouncements on the future of digital social research. Later on in this paper, we will return to these programmatic typifications of digital social research, but as our initial entry point into this area, we wish to focus instead on the technical and mundane practice and device of 'scraping' itself. As a technique of online data extraction, scraping seems to us of special interest, because it is an important part of what makes *practically* possible digital social research.

By way of context, it should however be noted that besides being a prominent technique, scraping is today also a notable news item. Scraping is granted special importance in narratives that are doing the rounds of publicity media about the analytic, normative and economic potential of digital social research. Newspapers from the New York Times to the Wall Street Journal, have recently run articles on scraping, making rather dramatic pronouncements about its social, economic, and epistemic implications. As a New York Times article reported, 'social scientists are trying to mine the vast resources of the Internet — Web searches and Twitter messages, Facebook and blog posts, the digital location trails generated by billions of cell phones. The most optimistic researchers believe that these storehouses of "big data" will for the first time reveal sociological laws of human behaviour — enabling them to predict political crises, revolutions and other forms of social and

economic instability.'[1] The Wall Street Journal chose to foreground the social angle, telling the story of a participant in online discussions on bipolar disorder, who found that his contributions had been scraped by a pharmaceutical company, or more precisely, a marketing firm working for a pharmaceutical company. "The market for personal data about Internet users is booming, and in the vanguard is the practice of "scraping." Firms offer to harvest online conversations and collect personal details from social-networking sites, résumé sites and online forums where people might discuss their lives."[2]

In this article, too, we want to explore the capacity of scraping to transform social research and to reconfigure the relations between subjects, objects, methods and devices of social research. However, rather than focusing on the 'human angle' foregrounded in the Wall Street Journal, or adopting the grand epistemological perspective of the New York Times, we want to focus on the *device* of scraping itself, and examine what kinds of social research practices it enables. Adopting such an approach enables us to examine digital social research *from the standpoint of its apparatus*. It is to adopt a more myopic focus on the concrete and everyday techniques and practices of online digital data capture. Such an approach offers a way to fill in anew a commitment long held dear by sociologists of science and technology, namely the insistence that science is best understood as a materially specific practice.[3] An investigation of scraping as a technique of data-collection opens up a perspective on digital social research *in-the-making*, as opposed to the declarations of intent and hopes for what digital social research might deliver as its final product as expressed in programmatic statements. Indeed, we will argue that digital social research might offer ways of renewing and radicalizing this commitment to research-as-process, as scraping may inform the development of 'real-time' or 'live' forms of social research, a term we take from Back and Lury

[1] Markoff,J. (2011) 'Government Aims to Build a "Data Eye in the Sky"', New York Times, October 10, https://www.nytimes.com/2011/10/11/science/11predict.html (accessed 11 January 2012). The stream of articles goes on: the Financial Times offered an account of a classroom at University College London, where students are devising ways to scrape social media to capture the most current financial information. Knight, S. (2012) 'School for Quants', *Financial Times*, March 2 (thanks to Patrice Riemens).

[2] Angwin, J. & Stecklow, S. (2010) '"Scrapers" Dig Deep for Data on Web', *Wallstreet Journal*, October 10.

[3] There have of course been several attempts to investigate the process of knowledge-making in the sociology of science and technology itself, including the rhetorical devices it deploys (Woolgar 1988). However, arguably these attempts were not so much concerned with the *technological* minutiae of doing social research.

(forthcoming).

We would like to propose that a more careful examination of online techniques of data capture may help to clarify some of the distinct analytic affordances of digital social research. Crucial in this respect is that scraping disturbs the distinction between the 'inside' and the 'outside' of social research. As we will discuss in the first section, the development of scraping as a data collection technique has been largely exogenous to social research. Scraping, then, is not native to social research and accordingly it risks to introduce 'alien' assumptions in this practice. Something similar applies, moreover, to the types of 'data' that scraping makes available for social research: distinctive about scraping as a data collection technique, we want to propose, is that it makes available *already formatted* data for social research. Scraping tends to involve the importation of categories into social research that are strictly speaking external to it. This circumstance, we want to argue here, is not necessarily a problem, or at least it may be turned into a virtue, and this insofar this circumstance can be rendered analytically productive for social research. It makes possible a form of 'live' social research, which approaches data formats as a source of social insight. Such an approach offers an alternative way of framing what scraping 'is for'. Instead of the live monitoring of live content, scrapers may also be deployed to analyse the *liveliness* of issues, a point we like illustrate through case study enabled by scraping. But first we need to talk about the scrapers themselves.

2. What is Scraping?

Formally speaking, scraping refers to a technical operation of Information Extraction, which is a field concerned with the automatic processing of data. Information extraction has acquired special saliency in the context of digitization and network media, in which the issue of information overload, the availability of overwhelming quantities of data, is 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 enabled by digital and networked media.[4] This promise is made usefully clear in another general definition of Information Extraction, which defines it as "the automatic extraction of *structured* information

[4] For example, in 2008 Google was processing twenty petabytes a day (Dean & Ghemawat 2008).

such as entities, relationships between entities, and attributes describing entities from *unstructured* sources" (Sarawagi 2007: 261; italics ours). Arguably, then, information extraction offers a solution to the problem of relevance, by deriving ordered data out of the unordered expanse of digital environments. And scraping refers a specific application of Information Extraction, as it provides a way of extracting specific fields or data elements not from individual data-bases, but from pages on the Internet.

As it is applied online, scraping makes it possible to bring together data from multiple locations, making the extracted data available for new uses, thereby enabling the 're-purposing' of online data.[5] One can scrape Web pages for images, or location data (town, region, country) or for 'keyword in context' data, which can then serve as a data set for research. This is also to say that the technique is easiest to use if the targeted content itself has structure. From an information extraction perspective, plain text is without structure. In another sense, however, scraping treats the Web blindly, as a limitless expanse from which only specific elements need to be brought in. Metaphorically speaking, one could say that scraping structures data collection as a 'distillation process,' which involves the culling of formatted data from a relatively opaque, under-defined ocean of available online materials. As such, scraping invokes the epistemic category of the 'plasma' usefully defined by Emmanuel Didier (2009; 2010) in his study of the invention of new social data collection techniques in New Deal America, like questionnaires. In Didier's study, the printing, circulation and collation of surveys made possible new types of expression of a relatively unformed 'panoply of elements.' Similarly, one could say scraping takes as its starting point the 'plasma' of online materials already in circulation, identifying specific data formats that have some currency in these

[5] There is some debate whether Web crawlers should be included among Web scrapers. A Web crawler - also known as an ant, a robot or a spider - "is a program or suite of programs that is capable of iteratively and automatically downloading Web pages, extracting URLs from their HTML and fetching them" and are used for a variety of purposes (Thelwall 2001). Web crawlers are most popularly known as one of the main components in the apparatus of Web search engines, where they assemble and index a collection of Web pages to allow users to query the index and return matching results (Page & Brin 1999). Related is the use of crawlers in Web archiving, for instance the Web crawler Heritrix developed and used by the Internet archive, where sets of Web pages are periodically collected and archived (Mohr et al 2004). A third use of crawling is for social research purposes. Crawling in this context is both used for large-scale data mining, where analytics is performed on Web pages or where they are analyzed for statistical properties (Thelwall 2001), and it is also used to demarcate and analyze more specific "Web spheres" or "issue networks" (Foot & Schneider 2002; Marres & Rogers 2005; Adamic & Glance 2005). While there is thus considerable overlap between what crawlers and scrapers do, in this paper we will limit our discussion somewhat artificially to scrapers, and the online capture of textual data in real-time.

messy, partly opaque streams, and then relying on these formats to extract analytically useful data for research.

More broadly speaking, however, scraping can be taken to refer not just to a technique but also to a specific *analytic practice.* Much scraping is done today in journalism, marketing, and policy research, and much of this activity concentrates on online platforms that offer live or 'real-time' data, such as Twitter. Scraping presupposes a socio-technical infrastructure, such as the increased availability on the Web of 'streams' and 'windows' that specifically address themselves to programmers and programmes, or to use the parlance, 'developers' and 'scripts'.[6] To put this differently, the popularity of scraping is closely connected with the rise of the so-called 'real-time Web,' which has been defined in terms of the equipment of Web services and online platforms for the provision of a continuous flow of fresh data (Berry 2011).[7] In large part because of its very freshness, this data is in need of constant disclosure and analysis. Scraping prepares these fresh online data for analysis, and a variety of tools and services are currently being developed to enable and accommodate practices of 'real-time' analysis in journalism, marketing, and policy-making. Figure 1 provides an example of the use of scraping in 'live' 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 expand and shrink in size depending on the frequency of their mentioning on the online chatter channel, thus providing an account of the scandal 'as-it-happened.'[8]

---

[6] This involves the provision and use of Application Programming Interfaces (APIs), which are currently being deployed and developed in a variety of practices and sectors. Scrapers stand in complex relations to API's - compared against the industry provided (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". See Watters, A. (2011) 'Scraping, cleaning, and selling big data: Infochimps execs discuss the challenges of data scraping', *O'Reilly Radar*, 11 May, http://radar.oreilly.com/2011/05/data-scraping-infochimps.html (accessed 12 January 2012).

[7] The rise of the real-time Web is closely related to technical developments of the Web (e.g. ajax, feeds, push notifications, APIs), and is seen most notably in Twitter's rise in usage and media popularity, Facebook's news feed and Google's increasing preference of freshness over relevance in search results. The question is, how realtime is realtime? (see Leggetter, P. (2011) 'Real-time Web or right-time Web?', *Programmable Web,* 17 March, http://blog.programmableweb.com/2011/03/17/real-time-web-or-right-time-web/ (accessed 12 January 2012).

[8] Richards, J., Graul, A., Shuttleworth, M., Santos, M., Dhaliwal, R., Stone, M.-L. & Dant, A. (2011) 'How Twitter tracked the News of the World scandal,' *The Guardian*, 13 July, http://www.guardian.co.uk/media/interactive/2011/jul/13/news-of-the-world-phone-hacking-twitter (accessed 12 January 2012). See also Rogers, S. (2011) 'What does Twitter think of the News of the World?', *The Guardian*, 7 July,

Figure 1: How Twitter tracked the News of the World scandal

Another example from the U.K. is ScraperWiki, which is a platform for developing and sharing scrapers. ScraperWiki is built both for programmers who want to make scrapers and for non-programmers who are after specific types of fresh online data, such as research journalists. Indeed, on its site ScraperWiki explicitly frames its services as including data provision for media, organizations and government. ScraperWiki also responds to a common problem with scrapers: their instability.[9] Scrapers are often custom-built, as they are designed to extract specific types of data from the Internet, and may also need to be adapted in response to changing access settings of the pages and sites to be scraped.[10] For this reason, many scrapers lead somewhat ephemeral existences, as they are taken in and out of use depending on needs arising, something which has been referred to as occasioning the rise of

http://www.guardian.co.uk/news/datablog/interactive/2011/jul/07/news-world-twitter (accessed 12 January 2012).

[9] As Helmond notes, the Web has a focus on freshness and as such "Internet methods are incessantly volatile due to the update culture of the Internet itself." Helmond, A. (2010) 'On the Evolution of Methods: Banditry and the Volatility of Methods', *Annehelmond.nl,* 17 May, http://www.annehelmond.nl/2010/05/17/on-the-evolution-of-methods-banditry-and-the-volatility-of-methods/ (accessed 12 January 2012).

[10] ScraperWiki also responds to another common problem with scrapers: as they are purpose-built, the code may easily get thrown away, break or is forgotten once it has been used, and so code and data is lost.

'plastic methods.'[11] In an ostensible effort to address this situation, ScraperWiki offers a Web-based, generally available platform that makes it more easy to use, develop, archive and manages scrapers. Finally, insofar as scraping involves the on-going development of tools and platforms, it arguably also involves a distinctive set of everyday practices, a culture, or, perhaps, forms of sociability.

Thus, ScraperWiki was one of the platforms introduced to students during a class in CAST, the Centre for Creative and Social Technology, recently established at Goldsmiths, University of London. Here, scraping brought with it a distinctive educational arrangement. In a 2-hour 'Sandbox' session, conducted in a computer lab, students were taught how to scrape through various ingenious pedagogical props, which included the projection of a 'recipe for scraping' on the wall (Figure 2). The 'recipe for action' detailed the relatively simple steps students were to follow, in order to extract a particular type of data from the online encyclopedia Wikipedia: a list of all the cities and towns were 'Occupy' protests where being held at that moment. The ScraperWiki platform itself enabled a particularly clever pedagogical trick, whereby each student could follow in their browser the scraping script being typed by the session leader in front of the class. In building our 'own' scrapers, we could either try to follow the steps in the recipe and type our own code. In case this didn't work, however, all we needed to do was to copy and paste the teacher's 'master code' into our own browser window, and voila we had build a scraper. This way or organising things was marvellous in giving us students a taste of what scraping feels like, but it also brings home the conceptual point already flagged above.

Scraping involves a series of steps in which formatted data is extracted out of an informational mess. To scrape is to build a chain from the relatively unformed mass of online data to formatted information, and along this chain relatively raw textual data is progressively stripped of its useless elements and formatted so as to produce a well-ordered, useable data set (something which is nicely captured in the programming concept of the 'pipe'). As such, scraping also nicely illustrates an argument from empiricist philosophy and sociology of science. Here, the process of inquiry does not consist of the projection of categories on a *fully-formed* empirical reality out there, but rather involves the creation of continuous chains which allow

[11] Martina Mertz in Helmond, A. (2010) 'On the Evolution of Methods: Banditry and the Volatility of Methods', *Annehelmond.nl*, 17 May, http://www.annehelmond.nl/2010/05/17/on-the-evolution-of-methods-banditry-and-the-volatility-of-methods/ (accessed 12 January 2012).

us to move from the relatively unformed to the formed, a process of extraction in which formats and their content emerge simultaneously (Latour 1999; Duhem 1954).

<insert Figure 2 and 3: Learning how to scrape with ScraperWiki in the CAST Lab, Goldsmiths (December 2011)>

3. Scraping as a device of social research

It is clear, then, that scraping is not a technique that is native to social scientific research. Rather, to use scrapers in social research is to import a technique into social research that is not native to it, and requires its adaptation, just as scrapers themselves are devices for re-purposing online data. Testifying to the precariousness of defining scrapers as techniques of social research is the fact that scraping is 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 online content – something of which Google among others is very critical – and these sites commonly use scrapers. Indeed, one could say that in some online discourses 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 recent algorithm updates, Panda, was specifically designed to combat aforementioned scraper sites.[12] Thus, it

[12] Wikipedia, (2012) 'Web scraping', Last updated 10 January, http://en.wikipedia.org/wiki/Web_scraping (accessed 11 January 2012).
Levy, S. (2011) 'TED 2011: The 'Panda' That Hates Farms: A Q&A With Google's Top Search Engineers', *Wired*, 3 March, http://www.wired.com/epicenter/2011/03/the-panda-that-hates-farms/all/1 (accessed 11 January 2011).

is as one set of application of scraping among many others, that social and cultural researchers are currently seeking to adapt scraping tools and services to their own purposes.

One example is info.extractor, a scraper developed by Chirag Shah and others for the extraction of public comments from the famous online social media platform Facebook for discourse analysis (Shah 2011). Another scraper, the Googlescraper – and less transparently, also known as The Lippmannian Device - is a tool for online textual analysis developed by the Digital Methods Initiative in Amsterdam.[13] This scraper scrapes the search engine Google, extracting parts of the query return pages from it (Figure 4). Arguably, the Googlescraper 'repurposes' the search engine Google as a 'research tool' (Brin and Page 1998; Rogers 2009; Weltevrede forthcoming). Providing a way to collect and visualise Google returns over time, the Googlesraper offers a means to study the rankings of Web sources produced by Google, and which are so influential in structuring traffic and 'attention' online. In another application of the Googlescraper, however, this tool also can be used by social and cultural researchers to perform textual analysis on the Web, as the scraper relies the search engine's index and search operators as a vehicle for conducting word frequency analysis in sets of Web pages. Using a device like the Google scraper, it then becomes clear that scraping raises potentially very tricky issues for social and cultural research. Using the Googlescraper, the question arises to what extent we are studying the device being scraped, in this case Google, or the forms of social life supposedly enabled by this device (Weltevrede forthcoming). Scrapers, that is, raise the question of how we establish the difference between *researching the medium* and *researching the social.* Just how far should we go in taking online 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 the 'object' of our analysis?

<insert Figure 4: Live view of the Googlescraper aka the Lippmannian Device>

As a technique of social research, however, scraping has some distinctive affordances, that seem to make these complications well-worth the trouble. First and

[13] Borra, E. et al, 'Googlescraper (Lippmannian device)', *Digital Methods Initiative* http://tools.issuecrawler.net/beta/scrapeGoogle/ (accessed 12 January 2012).

foremost, scraping solves a problem that social research shares with any other digital practices: it provides a solution to the circumstance that data out there on Web pages and platforms is not offered in a format that is at once usable. Scraping provides a way to aggregate data from disparate sources and platforms online. As such, scraping may appear to do no less than to unlock the 'sociological potential' of the Internet: it promises to make available for social research the very large quantities of user-generated data, that currently are being amassed through online platforms like Facebook, Twitter, Wikipedia and so on. Crucially, however, it should be noted that the very popularity of scraping is affecting these very opportunities, as more and more online platforms institute ways of regulating access to their otherwise 'generally accessible' platforms by bots and software agents. Thus, the chatter platform Twitter currently regulates access to its data through the Twitter-provided Application Programming Interface (API). At the moment of writing the Twitter API, for example, has a "theoretical maximum of 3,200 statuses" and this may be subjected to change at the companies will.[14] While social media platforms promise to make available a wealth of user-generated content or 'social data,' the way they format these data may end up placing severe constraints on social research. As dana boyd and Kate Crawford (2011) have flagged, a social research that relies on APIs risks rendering itself platform-dependent, and in effect accepts the blackboxing of its data collection methods (and more generally, the corporatization of its data).

Scraping however, holds a second attraction for social research, and this is that it may potentially solve the long-held research problem raised by online digital data, often referred to as a problem of 'dirty' data (Bollier 2010; Rogers forthcoming). Web data collection is often discussed in terms of the onerous and fraught task of having to process 'unstructured', 'messy', and 'tainted' data (Savage & Burrows 2007; Uprichard forthcoming). In addition to 'incomplete', Web data is frequently characterized as 'unordered' and 'dirty' (i.e. commercial). In relations to these issues, scraping provides a reminder that this issue is not just limited to social research. Indeed, as we have seen, Information Extraction, too, talks of Web data as unstructured data, but this field rather defines the cleaning and formatting of data as a research challenge. (Sarawagi 2007; Cafarella et al 2008). Indeed, scraping finds its very raison d'etre in the unstructured nature of online digital data. As such, the

[14] Twitter (2011) 'Things every developer should know', 11 July, https://dev.twitter.com/docs/things-every-developer-should-know (accessed 12 January 2012).

technique of scraping also suggests that it may be a mistake to consider 'quality' an attribute of data, rather than a methodological challenge. Scraping provides a reminder that the quality of data is in part of *processual accomplishment*: it partly depends on the operations that researchers perform on Web data, how good or clear these may become. In this respect, scraping may also be said to re-open the debate about the techniques that already used in social research for protecting and/or enhancing the quality of social data: what are currently the conventions and accepted standards for, for instance, anonymization and confidentiality of social data (Webber 2009; Savage & Burrow 2009; see also Gros 2012)?

If we understand the concern with the quality of online data to be relatively widespread, something important follows for our understanding of scraping as a social research technique. If the proliferation of devices like scrapers across social life are indicative of broadly shared efforts to address the problem of 'dirty data,' then social research is clearly 'not alone' in having to face this concern. Others are already looking for and finding solutions to this problem. This does not only mean that social research is likely to be scooped in its efforts to make online data amendable to research (Lynch 1991; Rogers 2009). Rather, we may have to define *the very field* of online data creation, management, disclosure and analysis in terms of on-going processes of data formatting and extraction. Social research, then, shares the challenge of how to extract tractable data with a heterogeneous set of other online 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 viewed as a vast collection of "completely uncontrolled heterogeneous documents" both in terms of internal variations and external meta information, defined as "information that can be inferred about a document, but is not contained within it" (Page et al 1998).[15] Indeed it was in this context that scrapers (and crawlers) emerged as devices capable of bringing order to - or extracting order from – the Web, as they make it possible to collect and re-structure large quantities of heterogeneous sources to be queried. Indeed, Google may be viewed as the uber crawling and scraping (or indexing) infrastructure, as it extracts hyperlinks and (anchor) texts and subsequently presents the data formatted differently, in a ranked

[15] Internal variations include language, vocabulary (e.g. email addresses, links, zip codes), type (text, image, sound) or format (HTML, PDF). Examples of external meta information include "reputation of the source, update frequency, quality, popularity or usage, and citations" (Page et al 1998).

list result page.[16] The prominence of these ordering devices online also raises the question of what is the distinctive contribution of scraper-enabled social research.

This, in our view, is what marks scraping as a 'device' of social research. If we approach digital social research from the standpoint of online data extraction, it becomes clear that *academic social science shares its research challenges and issues with a host of other actors*, *technologies and agencies that constitute the digital networked environment.* To adopt the standpoint of this 'device' on digital social research then has implications for how we understand the relations between the inside and outside of social research, and that between its objects and its methods. Firstly, it suggests a relatively broad definition of social research: Rather than envisioning social research as a practice that must be strictly demarcated from other forms of online data processing, we may approach social research as a relatively open-ended practice, which involves the deployment of a range of online devices and practices for the structuration of social data which themselves are not necessarily unique to social research. Elsewhere, Marres (2012) has referred to this phenomenon as the 're-distribution' of social research: taking up digital online tools, social research makes itself reliant on platforms, methods, devices for data processing, data formats, that have been developed in contexts and for purposes that are in many ways alien to those of social research. Social research, then, may ally itself with developing devices and practices for the structuration of social data. As a consequence, entities that are in some respects alien to the context of academic social research, nevertheless come to play a noticeable role in its organization, with various implications for analysis.

However, something else also follows from adopting a device-centred perspective on digital social research: it has implications for how we view the

[16] Initially Google's algorithm ranked purely by popularity measures, built around crawling and scraping hyperlinks and text around links, however increasingly Google is including other signals in their algorithm. Moreover, other factors are increasingly becoming privileged over popular reputational dynamics. Google states it uses over 200 signals to calculate the best results however, these signals can be clustered in a number of core types: Personalization by Web history, personalization by location, personalization by friends and freshness (Feuz et al 2011). These additional variables are however not always defining the results, but specific profiles are called upon depending on the type of query. Important in the context of real-time research, Google increasingly becomes a 'realtime' search engine for 'hot' or 'happening' issues. An important moment in terms of Google becoming realtime is their Caffeine update and secondly because they increasingly privilege fresh results over relevant results. Before the Caffeine update, their index was less fresh, i.e, the crawlers were sent out to index changes and new content on scheduled intervals, their main index being refreshed every week, whereas currently fresh content is updated in the index almost instantaneously. However behemoth, Google's data collection and analysis infrastructure is directed to collect specific data and to analyse and re-structure the collected data for a different purpose.

relation between the objects and methods of digital social research.[17] Indeed, it has been proposed that applying the concept of 'device' to social research is useful insofar as it highlights the relative *fluidity* of the distinction between object and method (Lury 2010; Law, Ruppert & Savage 2010; Marres 2011). To take up the sociological concept of the device, is to engage with a long and well-established tradition in the social studies of science and technology, and more specifically with the idea that if we approach research from the standpoint of its apparatus, certain customary distinctions become very hard to sustain, such as those between the techniques, methods, and objects of research (Duhem 1954; Latour & Woolgar 1979; Lynch 1991, Rheinberger 1997). If we consider social research devices like the 'social survey' or the 'focus group' in terms of the socio-technical settings and instruments involved in their conduct, it becomes very difficult to say where the methods of social research, their technical preconditions, and the phenomenon queried begin and end (Lezaun 2007; Law 2009). The object, the saying goes, is in part an artefact of the deployment of research devices, and in these devices technique and method are entangled beyond the point of repair.

Scraping provides a very fitting instantiation of this argument (as well as of other STS arguments, see Rieder and Rohle, 2012; boyd and Crawford, 2011; Yuwei Lin, 2012). In the case of scraping, too, it seems impossible to make straightforward distinctions between the instruments, methods and objects of digital social research. The Google scraper, for example, outputs its data as a structured text file (with categories) and in the form of a word frequency analysis (in this it is not unlike other scrapers). In this respect, it seems technically wrong to call this scraper a device for data extraction, as it also performs analysis and contributes towards the public presentation of results (in the form of so-called tag clouds) – a point to which we will return below. But more generally speaking, it seems a strange question to ask whether the Google scraper offers a technique or a method, as it so clearly operates on both levels. Or indeed, to ask whether the distributions of terms located by the Google scraper exist 'independently' from the devices deployed. Where in the ScraperWiki visuals in Figure 2 and 3 above are the method, where the technique, where the object? In practice, it seems impossible to disentangle these, and indeed, when we asked programmers questions about the difference between scraping as a

[17] The concept of the "technicity of content" introduced by Niederer and Van Dijck refers to the importance of including in analysis the technological tools and managerial dynamics that structure and maintain the content on Wikipedia, see Niederer and Van Dijck 2010.

method and a technique, they tended to give us very different answers. It is also to say that the distinction we made here above, between 'scraping the medium' and 'scraping the social' is probably best understood as a difference of degree: in some cases, online devices play an ostensibly large role in the structuration of data, while in other cases we can point to a discernable empirical object, which is not really reducible to the medium-architecture that enables it. In the case of scraper-enabled research, then, we may understand the boundary between object and method to be flexible, and to slide along a scale from 'all apparatus and no object', to 'part apparatus, part object'.

In some respects, it should not surprise us that these ideas about devices taken from the social studies of science and technology are so easy to apply to scraping. These STS arguments were partly the result of the importation of technical metaphors of information processing into the domain of the philosophy and sociology of science. Little wonder that they provide a useful language for making sense of the use of information processing techniques in social research! However, here we want to argue that we can derive some specific guidelines from this approach in terms of how to *deploy* informational devices in social research. We want to propose that it is *precisely* insofar scraping involves the importation of digital devices, data and data-formats into social research, that it may enhance the analytic capacities of digital sociology. As we mentioned, scraping formats data collection as a distillation process, as it provides a way of pulling specific fields or data elements from disparate pages, bringing together data from multiple locations, and extracting formatted data from socially available materials. As such, scraping takes advantage of the data formats that are implicit in social data, in order to structure these data. This state of affairs has distinctive affordances for social research, it seems to us, and may be *deliberately* deployed for analytic purposes.

This issue of 'pre-ordered data' has often been treated as posing epistemic, normative and political challenges for social research (Star & Bower 1999). Here we would like to propose that in the context of the rise of the aforementioned 'real-time Web', this circumstance has specific analytic affordances for sociology. To be sure, 'liveness' as a feature of social data, too, has been apprehended critically by some sociologists, who have flagged difficulties in terms of the limited life span and deterioration of 'live' data-sets (Uprichard forthcoming). However, one could say

that in this case scraping offers possibilities for turning analytic vices into virtues. The fact that data clearly have a life cycle online, might offer analytic opportunities for social research, especially in a context in which such 'life signals' may be rendered legible and analysable through scraping techniques. Similarly, then, to the question of the assumed structure or lack thereof of online data, the dynamism of these data may disturb certain assumptions of social research such as the assumed 'stability' of a high-quality data-set. These assumptions may be much less transferrable to online data environments than some sociologists seem to have instinctively assumed (Burrows & Savage 2009). Rather than seeing the liveness of data mainly as a problem (deterioration, incompleteness), we must then ask to what extend the live cycle of data itself may become a vehicle for analysis. In the remainder of this paper, we then want to outline a particular way of dealing with some of the difficulties associated with scraped data, by outlining a distinctive style of digital social research we call, following Les Back and Celia Lury, live social research.

4. Real-time research or the re-ordering of the empirical cycle

In order to clarify what is distinctive about live social research, it may be helpful to distinguish this approach from some prominent other definitions of digital social research. Indeed, as a 'medium-specific' technique of online data capture, scraping features prominently, though frequently implicitly, in current programmatic pronouncements and debates about digital social research. In recent years, several key terms have recently been proposed to capture the opportunities and challenges of digital social research: 'Big data' (Bollier 2010; Manovich 2011; boyd & Crawford 2011)[18] and 'digital methods' (Rogers 2009; forthcoming) are two of these terms designed to capture the newness and distinctiveness of the mode of social research enabled by what we referred to above as the 'second computational turn' in the social and cultural sciences. While not always called by name in programmatic writings associated with these key terms, scraping seems nevertheless central to them. Scraping is by no means the only, but an especially prominent technique through which 'big data' are delivered, and they have been key to the development of 'natively digital methods', or methods that are specific to digital media. However,

[18] Also see Anderson, C. (2008) 'The End of Theory, Will the Data Deluge Makes the Scientific Method Obsolete?', *Edge*, 30 June, http://www.edge.org/3rd_culture/anderson08/anderson08_index.html (accessed 12 January 2012).

these different key terms also imply rather different accounts of what is distinctive or innovative about scraping-enabled digital social research. After reviewing these two key words, we then propose a third, one which suggests to purposefully deploy a particular aspect of scraping in social analysis: the fact that it makes available pre-ordered data for social analysis.

To begin with the research programme that has received much popular attention, so-called 'big data' research focuses on very large data-sets and seek to push social research to the largest possible scale.[19] This research is often characterized as 'data-driven,' as the emphasis rests on the size and freshness of data-sets that become available for analysis using online tools of data capture. It also seems to be mainly exploratory in orientation, 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 of secondary importance as 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' (Douglass et al 2011; Manovich et al 2012), and other cultural analytics projects of the LA software studies group, was notably 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. The individual findings of the project – e.g. the distribution of visual styles among pages – seemed less spectacular as compared to these, in some sense, 'formal' markers of the analysis.

'Digital methods' research elaborates, 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 *devices* (Law, Savage & Rupert 2010; Rogers forthcoming). Here, it is the rise of digital platforms for data analysis, most notably search engines, that open 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. Cross-spherical analysis seeks analytical cohesion and simultaneously respects device-specificity by taking into account how information and issues are

[19] Big data refers technically to data that is too big to be processed by current tools and methods (Manovich 2011), 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 & Crawford 2011).

structured differently by different online engines and platforms - e.g. comparing the resonance of Issue Animals in the news, Web and blogosphere.[20] In so doing, digital methods research arguably opens up an alternative to 'big data' research, insofar as it foregrounds the opportunities that digital analytics offer for deriving significant findings from relatively *small* data-sets. Digital methods 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).[21]

'Real-time research' adds another key word to the debate about digital social research, and arguably points towards yet another type of social research enabled by scraping, although the concept in and off itself does not strictly exclude the previously mentioned approaches. This term has been proposed by Back, Lury and Zimmer to characterize the transformation of the spaces and times of social research in the context of digital culture: the increasing valorisation of instantaneity and liveness, the drive towards the condensation of the past, present and future in social research, or the conjuring up of an 'eternal now' (see on this point also Uprichard, forthcoming). But here we would like to suggest that 'real-time research' highlights a third way in which the digital is refiguring empirical social research. The digitization of social life, and the method of scraping more in particular, signals a re-ordering not just of the times and spaces of social research, but also *of the empirical cycle* itself. Scraping, as we noted above, tends to deliver 'pre-ordered data', and in the process of scraping, the collection and analysis of data seem to be very much entangled. Talking with programmers it quickly became clear that scrapers do not just collect data but analyse them in one and the same go, as they 'parse' the data culled from the Web. And as we already noted, to extract data scrapers rely on information structures embedded in the media, such as the 'hashtags' used to identify twitter posts. We might therefore say that in scraping-enabled forms of social research, analysis 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 'meta-data' and thus treated as instrumental in operationalizing analysis? To highlight the latter possibility is to go against a particular assumption of debates

[20] Niederer, S. & Weltevrede, E. (2008) 'Issue Animals Research', *Digital Methods Initiative*, https://wiki.digitalmethods.net/Dmi/IssueImageAnalysis (accessed 12 January 2012)

[21] 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.

about digital data in sociology: the aforementioned suggestion that digital online data-sets must be characterized as 'unordered'. This assumption seems to be an artefact of the particular vantage-point from which sociologists tend to approach digital social research, namely as a form of research that is to be compared to experimentally controlled social research such as sample-based survey research and the analysis of textual corpora (Savage & Burrows 2007; see also Newman, Barabási & Watts 2007). In comparison with complete databases of 'single source' survey or interview data, online data-sets may seem unordered, unruly, incomplete, unsystematic. From the standpoint of the digital data deluge, however, the opposite seems true: because of the size and freshness of online data-sets, digital research must rely on highly specific markers present in the data (or 'ordering devices') like links, rankings, date stamps, and hashtags in order to gain traction on data (see on this point also Fuller 2008). We propose to define as 'live,' social research that seeks to derive its analytic capacities from the 'pre-formatting' that is distinctive of online social data.

5. Different digital empiricisms

To sum up, different types - or typifications - of digital social research can be said to assume and deploy different *empirical imaginaries* (Law 2004). Thus, the empiricism of 'big data' can be opposed to that of 'real-time research'. Arguably, 'Big data' can be said to re-instate a positivist version of empirical research, or even, to enact a 'revenge of modern empiricism', as associated with the large-scale programmes of quantitative social science associated with sample-based survey analysis that rose to prominence in the post-war period in American social science in particular (Burrows et al 2008). Insofar as big data research is data-driven it also invokes a central principle of classic modern empiricism, namely the notion that data-collection precedes theory-formation. This type of research, furthermore, also re-instates the division between data and their interpretation, as it frames research in terms of operations of 'pattern-recognition' on large, unordered data-sets. (This division is sometimes re-enforced by critics of big data research, which criticize it for prioritizing data-collection over theory formation, and warn that analysis and theory risk to become an 'after-thought' in this type of research.)[22]

[22] Big data research also tends to adhere to classic empiricist notions of experimental control (Newman et al 2007), the idea that experimental research ideally requires maximum if not total control over the variables under scrutiny (Marres forthcoming)

By contrast, real-time research rather aligns itself with the second empiricism (or post-empiricism) associated with philosophers like Pierre Duhem, W.C. Quine, Hans-Jorg Rheinberger, Isabelle Stengers and Bruno Latour. Real-time research highlights the *formatted* nature of much digital data – the ways in which digital data is networked, ranked and tagged – and as such, it can be said to assume the 'theory-ladenesss' of observation. On a general level, it highlights the circumstance that data, whenever we encounter it in any tangible form, is already marked by conceptual categories, and that, for this reason, there is no such thing as 'raw data' (see Gros 2012 for an elaboration of this point).[23] This kind of approach draws on holistic theories of knowledge, which have long suggested that, in practice, data and analysis cannot be distinguished in any easy or straightforward way. With digital social research, this insight becomes ostensibly relevant to *the methods of social research itself.*[24] Here, the idea that 'there is no such thing as raw data' and the claim that empirical technologies format the phenomena they purport only to measure, become very much applicable to the research techniques deployed by sociologists themselves: to their tools of textual and network analysis, not to mention data collection techniques like crawling and scraping. Indeed, digital research techniques, like scraping, make it possible to render this insight – data are always already formatted - analytically useful.

In the remainder of this paper we would like to specify some empirical uses of 'the information format'. However, one important consequence also follows from the above for wider debates in digital social research about the differences between 'data-driven' versus 'interpretative' social research, and their relative merits, in a digital age. These debates are frequently framed as the 'next chapter' in a long-standing 20th century intellectual stand-off that pitched positivism against phenomenology, facts against interpretation, and quantitative methods versus qualitative research. However, if we characterize digital social research in terms of the re-ordering of empirical research – in which analysis now precedes or coincides

[23] This insight fueled the post-empiricist philosophy and sociology of science during the second half of the 20th century. Thus work suggests that the distinction between data, experimental device, hypothesis, and theoretical postulates is an unstable and precarious one, one that may easily break down and requires continuous work to be maintained. This in turn led to the adoption of a 'holistic' view of experimental research, according to which the distinction between data, theoretical postulates and experimental setting is an unstable and precarious one, that can only be practically accomplished (Duhem 1954; Quine 1951).

[24] This claim has also informed 'performatist' studies of science and technology, which seek to demonstrate how 'seemingly neutral' empirical devices, like a stock index or a weather station, in effect frame and format the phenomenon they purport merely to measure.

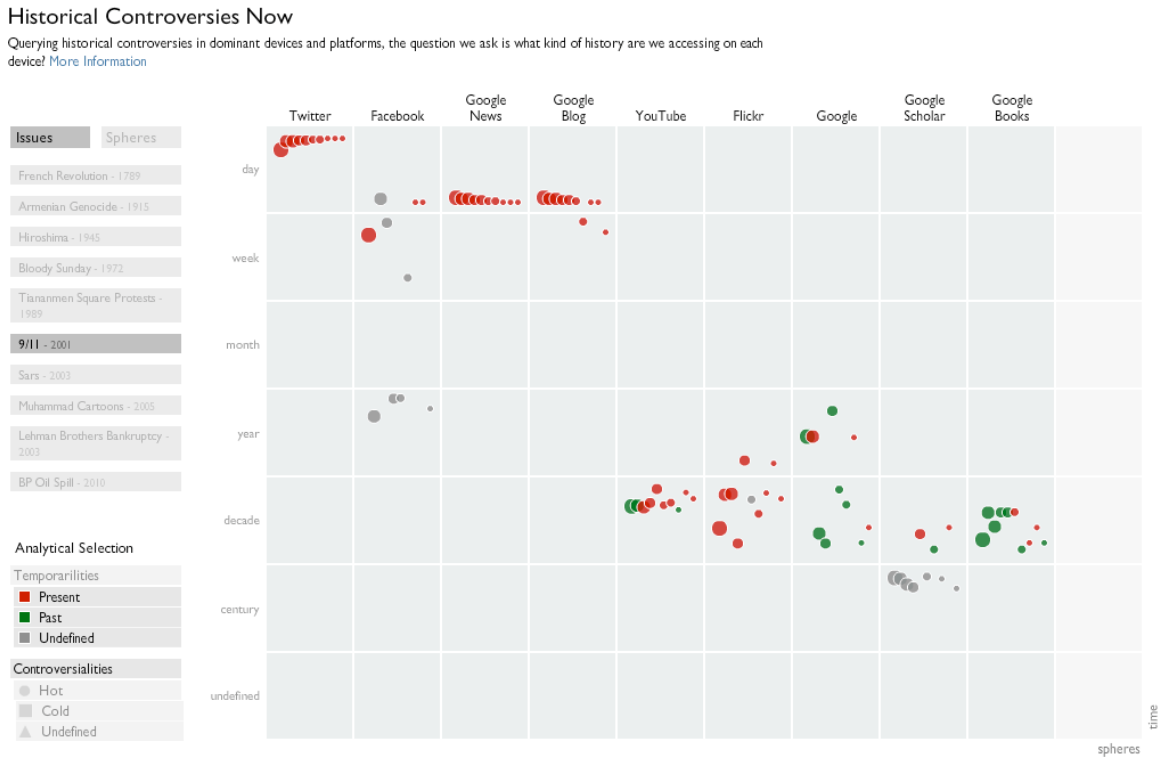
with data-capture, and in any case cannot be strictly speaking distinguished from it – then it is rather the *fusing* of these two categories in digital social research that we must come to terms with (see on this point also Law, Savage & Ruppert 2010; Latour 2012). To put it grandly, it is the *collapse* and not the re-instatement of the established dichotomy between data-driven and interpretative research - which has done so much to provide debates about the role of the humanities and social science with their normative direction - that we must come to terms with in the context of digital social research.

6. Live sociology: Meta data as the new social data?

Rather than follow through these implications on a general and abstract level, however, here we would like to flag a specific methodological consequence of doing social research with the type of formatted data that scrapers help to extract. Scraping enables a style of 'live' social research, in which 'extraneous' data features like the absence or presence of a source or term, the frequency of their occurrence, and other source attributes like freshness or connectedness, provide key indicators. These data-attributes could be called meta-data, even if they are conceptually and distinct from (but closely related to) technical meta-data features, such as name, size, data type. Meta-data are of course data themselves, but they are also descriptive of, and structure, other data. We call this type of research, following Back and Lury, 'live' social research, as it seeks to 1) draw on media streams to extract categories or data formats from media practices for social research, instead of stripping data from its formats and imposing formatting on the data from outside, and 2) it relies on these embedded data formats to identify specific life signals - like frequency of mentioning, absence/presence and data-stamps - to track the dynamic composition of social life and its issues online.

Such a form of life social research then values positively the fact that the information extraction techniques built into online devices, *pre-format* data for social and cultural research. As we saw above, search engines format heterogeneous data-sets by collecting and processing data through digital makers such as hyperlinks and anchor texts, and online platforms pre-format social transactions in software and interface design. In both cases the online devices already order, or more precisely, pre-format, the data for social research. However rather than treating these formatting effects as something than contaminates our data with noise or

negative bias, we propose that social research may partly derive its analytic capacities from these effects. That is, real-time social research may ask how data formats that are already implicit in the data, may be analytically deployed to structure social research and generate 'findings'. Indeed, it is to the extent that online digital data are already pre-formatted with the aid of these ordering devices, that it is possible to do small research with big data: these markers are like levers with which to leverage large data. An example of this type of approach can be found in the project Historical Controversies Now, which compares ten recent and less recent historical controversies. [25] The project tests the hypothesis that online digital media collapse the present and the past into an algorithmic "nowness", showing how different online platforms formats the space-times of controversy differently. For example, Twitter search organizes maximum 140 character tweets by freshness (i.e. date stamps) providing brief and happening re-workings of past controversies, whereas Google Scholar, which organizes scholarly writings by academic relevance (i.e. among others citations), provides well-cited academic contemplations of these same controversies (See Figure 5).



<insert Figure 5: Historical Controversies Now >

[25] Weltevrede, E., Rooze, M., Poell, T.,Feuz, M., & Dagdelen, D. (2010) 'Historical Controversies Now', *Digital Methods Initiative*, https://files.digitalmethods.net/var/historicalcontroversies/ (accessed 12 January 2012).

Importantly, as 'live' social research seeks to take advantage of, and aligns itself with, data formats that are inherent to the medium, it offers a particular way of dealing with the epistemic issues raised by scraping and digital social research more widely. As we have seen, a host of characteristics have been attributed to digital social data, which are widely perceived to problematize its analytic affordances and its 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, platform-dependency (data-centrism), and the blackboxing of method. Thus, dana boyd and Kate Crawford (2011) have drawn attention to the relative obscurity and general dissapointingness of the 'data hoses' offered by platforms like Twitter and Facebook, noting that what gets presented as a big hose offers little more than a sprinkle, in what also amounts to a nice commentary on the inbuilt gender biases in these debates. Others have, in an equally pleasurable ways, 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.[26] We would like to suggest that one of the advantages of adopting the approach of live social research is that it entails a change in status of this type of epistemic trouble.[27]

For social research that focuses on meta-data, the structure and dynamism of online data must be primarily regarded in generative terms, rather than in the negative terms of a 'bias' that prevents us from seeing clearly. Here data is amenable to analysis precisely to the extend that it is ranked, linked, stamped and tagged, and dynamically so. More generally speaking, this means that from the standpoint of real-time research, the epistemic trouble with digital social research - platform-dependency, data-centrism, blackboxism – do not necessarily discredit

[26] Slee, T. (2011) 'Internet-Centrism 3 (of 3): Tweeting the Revolution (and Conflict of Interest)', 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#paper2 (accessed 12 January 2012)

[27] Especially when sailing on the compass of the positivism/phenomenology opposition, these epistemic issues feature negatively: they appear as so many reasons why empirical social research fails to deliver on its promises. However, once we recognise the complication of this opposition in digital social research another strategy may be conceived: rather than interpret these problems negatively, we should offer an alternative valuation of it: they do indeed problematize empirical social research, but, as such, these problems offers occasions for inventive research and they invite us to re-imagine the relation between the problems of digital sociology, and the problems of the digital society.

digital social research. In real-time research, one could say, these issues do not threaten social research 'as if from the outside', but trouble it in a much more immanent fashion, and may also become an object of social research. By plotting the distribution of returns, the above case study Historical Controversies Now, for instance, turned platform-dependency – the ways in which online platforms structure the space-times of inquiry – into a topic of investigation. This is also to say, the epistemic issues associated with digital social research noted above are relevant well beyond the confines of 'rigorous social and cultural science': indeed, they closely resemble many of the issues that trouble digital social life, much more generally: data-centrism, platform-dependency, the opacity of knowledge technologies, these are not just the problems of social research, but of the information society at large. The epistemic trouble generated by scraping are then not just the problems of social and cultural science, but rather that of many social practices that involve the collection, management, analysis, and operation of digital social data by these means. In this context, the 'best' attitude to these epistemic trouble raised by digital social research 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, and to make their effects visible and reportable for practical purposes.

7. Conclusion: from live media to the liveliness of issues

Live social research, then, is social research that seeks to render analytically productive the formatted, dynamic character of digital networked data - it is a research practice that seeks to render explicit the instability or dynamism or 'shape-shifting' of online data, turning these into a resource and an object of digital social research. In this respect, however, one crucial question is how 'live' social research relates to, and is different from, the research practices associated with the 'real-time Web,' which we discussed above. The latter type of research, we said, deploys scraping tools in order to capture fresh data about current themes, sources and actors. What makes the type of live social research under discussion here different is that it adapts techniques of online data extraction to determine not just the 'liveness' of specific terms – how prominent are they in current reporting? - but their *liveliness*.[28] From this standpoint, the key question is not what topics, sources and

[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 & Rogers 2005).

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 *happening*: which terms, sources, actors are the most active, which fluctuate most interestingly over a certain period? (Rogers 2002; Marres 2012). We would like to conclude this paper by outlining a case study that could help to further specify this difference between researching liveness and liveliness. (We realize it is a bit odd to end our paper with an empirical case, and thus, have analysis precede data-collection here too, but this was how the paper developed.)

In order to determine how to establish this distinction technically, methodologically and analytically, we decided to scrape two online devices, Twitter and Google, for a small number of terms, which were likely – we thought - to display both dynamics, liveness and liveliness. With the help of programmers affiliated with the Digital Methods Initiative, we scraped Google and Twitter for two key words 'austerity' and 'crisis', for the duration of some months.[29] As part of this scraping exercise, data on the relative prominence (currency) of these key terms was collected as a matter of course, but we were especially interested in establishing the *variability* of these two key words (how happening are they?). Thus, we relied on scraping techniques in order to monitor which language is associated with our two key-words in these devices, with the aim of identifying fluctuations in and of this language, or the 'dynamic composition' of the issue. Thus, we asked, are 'austerity' and 'crisis', in a given media stream, 'lively' issues? Are we able to distinguish 'newsy' and more 'social' forms of variation in this issue's composition? Do newsy issues, for instance, come in bursts, traceable to newsworthy events? What would be the signature of issues that are also socially alive? Such a analysis turns the 'instability' of online data-sets into an object of analysis, treating variations in the key-words with which our term is associated (its co-words) into an indicator of the liveliness of certain terms, or perhaps their 'state of issuefication'.

In adopting this method, we decided to rely on techniques of co-word analysis, which is well-established in scientometric research and other forms of content analysis (Leydesdorff & Welbers 2011; Callon et al 1983; Danowski 2009; see also Marres 2012). Importantly, co-word analysis as a methodology has traditionally relied on the formatting of the medium for the structuration of social data. Thus,

[29] We are grateful to Erik Borra and Bernhard Rieder for their help. Using the Twitter and Google analytics platforms currently under development, we scheduled an ongoing data collection from 1 January in Google and Twitter for 'austerity' and 'crisis.'

Callon et al (1983) rely on the keywords used to index academic articles in order to establish its lexicon of co-words.[30] This can be contrasted with some recent adaptations of co-word analysis, such as the tool Wordij, which analysis a wide array of textual sources including news stories, blogs, discussion forums, IRC chat, open ended survey responses and emails (Danowski 2009). Different from Danowski and akin to early versions of co-word analysis, in our version we chose the tweet and the snippet as unit of analysis for the co-occurrence of words. In analysing Twitter and Google data, we thus relied on the information formats that are central to the operation of these online devices, as our principal units of analysis: the tweet in the case of Twitter and the 'snippet' for Google, a title and string of key-words that Google returns for each individual page in its query return list.[31] These two formats structured our data insofar as we limited our co-word analysis to the identification of associated language within the tweet or snippet, in order to locate relevant fluctuations in our data. Our space of analysis thus delineated, we subsequently focused on two types of what we called here above meta-data: absence and presence of co-words, and secondly, the date stamp, which is also slightly different per Web device,[32] but which in both cases enabled us to work with the fluctuation of words over time.

One of the major challenges in declining to research liveness, and deciding to focus on liveliness instead, we soon found out, is that we loose a powerful instrument for data reduction. To ask which term is the freshest, or which is the most popular, we were forcefully reminded, is a splendidly easy way of boiling down vast

[30] Wordij uses a sliding window to demarcate the proximity radius of co-words: "The approach can be thought of as using a window, 'n' word positions wide, that slides over the text, identifying, counting, and aggregating all word pairs that appear within the window" (Danowski 2009).

[31] Web devices Google and Twitter each design access to the data differently and thus the scraping procedure slightly varies accordingly. In Google we collected a daily snapshot of the result page for 'crisis' and 'austerity,' and in Twitter we collected all tweets containing the keywords coming through the public API *as it happens*. The unit of analysis from the Google result page is the individual result – set to 100 results per page – which consist of a title and a snippet; in Twitter the unit of analysis is the individual tweet. For instance in one of the cases below we used 4 days of data, which contains approximately 400 snippets and just under 145.000 tweets. Note that for analytical purposes we chose to retain only unique tweets, not taking into account retweets. The use of retweets as analytical device is useful to boil down the number of tweets following a popularity metrics, or put differently, to find those tweets that have pass along value (Weltevrede 2011). In this case the aim is instead to filter out words following a liveliness metrics, focusing on the intensity of connections between words contained across a heterogeneous set of tweets (Marres 2011).

[32] Whereas Twitter provides each individual tweet with a date stamp indicating when it was published, Google only returns a date stamp for relatively small portion of results. The date stamp we collected for the Google snippets is instead derived from the day we captured the result page and the result is found present in the top 100 results for the query 'crisis'.

amounts of data to a few significant words, something which our colleague Emma Uprichard, had long warned us about, in flagging the issue of 'data reduction'. If it is not the currency of terms of seek to establish, but their liveliness, how do we decide which fluctuations are relevant, in a methodologically sound manner, among the 20.000 word associations we for instance discovered in four days of tweets? One way would be to devise a criterion of variability: which words are most variable in their composition (co-words)? Which of our issues are associated or 'charged' with many different co-words on different platforms, or in different 'media spheres'? For instance, is crisis an economic issue in some places, both mostly environmental in others? However, with this feature currently lacking in our co-word machine in-the-making, however, we decide to re-admit a measure of currency, allowing the measure of frequency of mentioning to inform our analysis of liveness.[33] Thus, in this initial sketch of the liveliness of 'austerity' on Twitter and Google, we only included co-words that occurred a minimum number of times in our corpus (Figure 6 and 7).[34]

<insert Figure 6. Co-words related to austerity derived from Google result pages, 1-31 January 2012.>

<insert Figure 7. Co-words related to austerity derived from the Twitter stream, 1-31 January 2012.>

In exploring our initial co-word maps, using the network visualisation and analysis software package Gephi, it also became apparent that the hold of 'liveness' on our respective platforms, Twitter and Google, goes well beyond the prevalence of currency measures in scraper-enabled analyses of these platforms. As it turned out, the very content of our analysis reverberated with the 'language of the wire' – the terms that had been occupying the news of the period in question (greece, imf, bailout, protest). (Thus, the co-word network graph based on Google snippets

[33] Another way would be to devise a criterion of variability: which terms are associated with a varying set of co-words, with X numbers of co-words varying in period X.
[34] Co-words in the month January in Twitter where only retained if they occur in 100+ tweets per day, and in Google we retained those ties that occur in 30+ snippets per day. Visualizing the co-word networks in Gephi, the connecting word 'austerity' was removed to allow a sparser graph. Additionally also an editorial choice was made to remove formatting language specific to the medium – e.g. 'retweet,' 'via,' 'hours ago.'

(Figure 6) shows different national governments and their respective austerity measures – cuts in Romania are accompanied by protests; a new package for Italy and Spain; and Athens' parliament approves new bailout.[35] In many cases, the prevalence of the news in these channels was literally apparent: in the Twitter co-word network clusters formed around specific documents that receive attention, such as a guardian article with the title "IMF warns of threat to global economies posed by austerity drives."[36] For both platforms, we checked which sources are referenced in the data: a host-count of sites found in the Google results and tweets containing the word pair austerity and crisis indicates that both spaces are dominated by news sources. Austerity, then, is a highly newsy term, both in terms of the terms and actors composing this issue, at least in the online platforms under scrutiny.

In the second sketch of the liveliness of 'crisis' on Twitter and Google, we decided to only use hashtags in Twitter instead of co-words and thus resorting to medium specific data formats. This methodological choice leads us to include a reasonably sized set of connected hashtags with a minimum strength of connection of 5. Additionally, we decided to reduce the data by focusing on a limited set of 4 days of data instead of excluding fine-grained co-words from analysis by raising the threshold of connection strength (Figure 8 and 9). Interestingly, by focusing our analysis on co-hashtags the lively terms emerge around newsy terms: wecanbeheros, screwyouassad, freechina en eurodebtcrisis.[37] The visualisation shows static terms – i.e. terms that occur across four days – in dark grey, and terms that are dynamic – i.e. occurring in three days or less – in red. The static terms provide the context for the dynamic words.

The difference between the two Web devices for the same term is again significant. Compared to the Twitter co-words, which feature the euro crisis and in the periphery figure specific local crises, the co-words taken from the Google results focus more on the infrastructure of crisis – support, help, services, support, rape,

[35] The co-word network also shows fears that the euro crisis and specifically the austerity measures might be killing the European economy. Additionally, there is a cluster with definitions of the word austerity. In the Twitter stream the countries figure less prominent (Figure 7). The Romanian protests and the new measures in Greece are present, but more prominence is given to the IMF and the euro crisis. It is notable that specific names figure more prominently, such as Joseph Stiglitz, Jeff Madrick and Paul Krugman- all three are economics critical of austerity measures.
[36] Elliot, L. (2012) 'IMF warns of threat to global economies posed by austerity drives', *The Guardian*, January 20. Available at: http://www.guardian.co.uk/business/2012/jan/20/austerity-warning-international-monetary-fund (accessed 20 March 2012).
[37] Due to the multilingual term 'crisis' there are clusters of different languages to be found, including Spanish, Dutch and English.

women. The comparative co-word analysis thus shows the multiplicity of the object in different Web devices. Furthermore, the co-word visualization from the Google results seems to indicate that Google results are less volatile, however, this is most likely due to the limited number of days under analysis. The project and tool Issuedramaturg convincingly captured the volatility of Google results over a longer date range.[38]

<insert Figure 8: Issue profiling crisis with related hashtags in Tweets, 23-26 January 2012.>

<insert Figure 9: Issue profiling crisis related words in Google results, 23-26 January 2012.>

The challenge of adapting techniques of scraping to the purposes of social research, and to move from the analysis of liveness to liveliness, is thus not just a methodological challenge, but equally a challenge addressed to 'the medium' itself. Importantly, though, when we delved into the detail of the newsy 'austerity' co-word data returned by our scraper-in-the-making, we did seem to find clusters of terms, which to our minds at least reverberated with the social: for instance, valentinespoems; merkelnotmychancellor; solidaritywithgreece; bankstergansters; inflation; depressed; isolated; capitalism; silver; assholes.

Here we were reminded of a warning made by sociologists inspired by Gabriel Tarde, which has raised some serious doubts about the sociological ambition to move 'beyond' journalism: the circulation of news, they argue, plays a central role in structuring social life in mediatized societies, and if we were to try and edit it out of our analysis, we would end up turning much of social life into a mystery (Latour & L'Epinay 2008; Barry 2005). In adapting scraping for the purposes social research, it would then be a mistake to try and 'route around' the news. Rather, our challenge is that of how to render legible the much more fine-grained variations we are intuitively inclined to associate with the social: cruelly or perhaps justly so, we found that the more 'social' – less regulated, less official, less formal - co-word variations were much more resistant to visualisation than the broad-stroked newsy

[38] The project is located at: https://dramaturg.issuecrawler.net/issuedramaturg_story/911truth.org.html, accessed 22 March 2012.

dynamics. The challenge of adapting scraping as a technique of social research is then at least in part a challenge of data selection and visualisation.

In future research, it would be interesting to investigate other issues that are less 'newsy' – for example, climate action and community energy, which equally figure prominently in policy documents and publicity by environmental and social organisations. It would also be interesting to investigate further the various paces of Web devices - such as in Historical Controversies Now example - and to explore how patterns of liveliness materialise differently in these algorithmic dynamics of Web platforms and devices. This, indeed, is indispensable, if we are to address what we now regard as the central question that that scraping online platforms opens up for social research, which is that of how to characterize different modes, styles and forms of the 'happening' of issues. The question that is, cannot not just be the general one: which issues are most actively fluctuating? Rather, as it delivers masses of data detailing more or less relevant fluctuations, scraping online platforms forces us to become more precise as to what type of term fluctuations we are after. This type of 'data overload' does not, however, seem to us merely an artefact of online digital media. Rather, it forces us to offer a more differentiated analysis of a process that we have referred to as issuefication, and is also referred to as the 'heating up' of issues (Callon 2007). In ANT-inspired controversy analysis, this dynamic has too often been minimally or rather generically defined as a process of 'politicization' or 'unblackboxing' and operationalized to the rather broad-stroked indicator of an increase in information production and expansion of circulations about a given issue. But of course, objects may 'heat up' in very different ways; it is not just that some issues that are lively or hot in some places may be cold or dead in others. They may heat up in different ways, say more newsy or more expertly or – dare we say it - 'social' ways?

Acknowledgements:
The authors wish to thank participants in the Digital Methods Winter School for some very helpful comments on this paper, especially Carolin Gerlitz, Erik Borra and Bernhard Rieder.

References

Adamic, L. (2005) 'The Political Blogosphere and the 2004 U.S. Election: Divided They Blog', *Proceedings of the 3rd international workshop*, pp.1-16. Available at: http://dl.acm.org/citation.cfm?id=1134277 [Accessed January 8, 2012].

Adkins, L. & Lury, C. (2009) 'Introduction: What Is the Empirical?', *European Journal of Social Theory*, vol. 12, no. 1, pp.5-20. Available at: http://est.sagepub.com/cgi/doi/10.1177/1368431008099641 [Accessed August 15, 2011].

Back, L. (2010) 'Broken Devices and New Opportunities: Re-imagining the tools of Qualitative Research', *ESRC National Centre for Research Methods NCRM Working Paper Series*, November.

Back, L. (forthcoming) 'Introduction', *The Sociological Review*, Special issue on Live Sociology, eds L. Back & N Puwar.

Barry. A. (2005) 'Events that Matter', *Paper prepared for the workshop on Gabriel Tarde*, University of London Senate House, London, 1 December.

Berry, D. (2011) 'Real-Time Streams', In *The Philosophy of Software: Code and Mediation in the Digital Age*, Palgrave Macmillan, New York, pp. 142-171.

Bollier, D. (2010) 'The Promise and Peril of Big Data', *The Aspen Institute,* Washington, DC. Available at: http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf [Accessed August 15, 2011].

Bowker, G.C. & Star, S.L. (1999) *Sorting things out: classification and its consequences*, MIT Press, Cambridge MA.

boyd, d. & Crawford, K. (2011) 'Six Provocations for Big Data', *Paper presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society,"* pp.1-17.

Brin, S. & Page, L. (1998) 'The anatomy of a large-scale hypertextual Web search engine', *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp.107-117.

Callon, M. (2009) 'Civilizing markets: Carbon trading between in vitro and in vivo experiments', *Accounting, Organizations and Society*, vol. 34, pp. 535-548.

Callon, M. & Courtial, J., Turner W. & Bauin S. (1983) 'From translations to problematic networks: An introduction to co-word analysis', *Social Science Information*, vol. 22, pp. 191 – 235

Cafarella, M., Madhavan, J. & Halevy, A. (2008) 'Web-scale extraction of structured data', *ACM SIGMOD Record*, vol. 37, no. 4. Available at: http://dl.acm.org/citation.cfm?id=1519112 [Accessed January 10, 2012].

Carnap, R. (1966) 'The Experimental Method', In *An introduction to the philosophy of science*, ed. M. Gardner, Basic Books, New York, pp. 40-47.

Danowski, J.A. (2009). 'Network analysis of message content', In *The content analysis reader,* eds K. Krippendorff & M. Bock, Sage Publications, Thousand Oaks, pp. 421-430.

Dean, J. & Ghemawat, S. (2008) 'MapReduce: Simplified Data Processing on Large Clusters', *Communications of the ACM*, vol. 51, no. 1, pp.107-113.

Didier, E. (2010) 'A theory of social consistency', *paper presented at After Markets*. Said Business School, University of Oxford, Oxford.

Didier, E. (2009) *En quoi consiste l'Amérique: Les statistiques, le new deal et la démocratie*, La Decouverte, Paris.

Douglass, J., Huber, W. & Manovich, L (2011) 'Understanding scanlation: How to read one million fan-translated manga pages', *Image and Narrative,* vol. 12, no. 1, pp. 190–228.

Duhem, P.M.M. (1954) 'Physical Theory and Experiment', In *The aim and structure of physical theory*. Princeton University Press, Princeton, pp. 180-218.

Feuz, M., Fuller, M. & Stalder, F. (2011). 'Personal Web searching in the age of semantic capitalism: Diagnosing the mechanisms of personalization', *First Monday*, vol. 16, no. 2. Available at: http://

firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/ view/3344/2766> [accessed 9 September 2011].

Foot, K.A. & Schneider, S.M. (2002) 'Online Action in Campaign 2000: An Exploratory Analysis of the U.S. Political Web Sphere', *Journal of broadcasting & electronic media*, vol. 46, no. 2, pp.222–244. Available at: http://www.tandfonline.com/doi/abs/10.1207/s15506878jobem4602_4 [Accessed January 7, 2012].

Fuller, M. ed. (2008) *Software Studies: A Lexicon*, MIT Press, Cambridge, MA.

Gros, A. (Forthcoming) 'Overflowing Personhood: Personification, Detachment and Anonymisation and the Shifting Value of Personal Data', *The Sociological Review,* Special issue on Measure and value, eds L. Adkins & C. Lury

Latour, B., Jensen, P., Venturini, T., Grauwin, S. & Boullier, D. (forthcoming)'The Whole is Always Smaller Than Its Parts: A Digital Test of Gabriel Tarde's Monads', *British Journal of Sociology*.

Latour, B. (2009) 'Bruno Latour on Mapping Controversies: An Introduction to the MACOSPOL project', *Mappingcontroversies.net,* Available at: http://mappingcontroversies.net/Home/PlatformMappingControversiesVideoIntroduction [accessed January 7, 2012].

Latour, B (1999). *Pandora's Hope. Essays on the Reality of Science Studies*. Harvard University Press, Cambridge, MA & London.

Latour, B. & Woolgar, S. (1979) *Laboratory life: The construction of scientific facts*, Princeton University Press, Princeton.

Law, J. (2009) 'Seeing Like a Survey', *Cultural Sociology*, vol. 3, no. 2, pp.239-256. Available at: http://cus.sagepub.com/cgi/doi/10.1177/1749975509105533 [Accessed August 5, 2011].

Law, John (2004) *After Method: Mess in Social Science Research*, Routledge, London & New York.

Law, J., Ruppert, E., & Savage, M. (2011) 'The Double Social Life of Method', *CRESC Working Paper Series,* vol. 95, pp. 1-11. Available at: http://heterogeneities.net/publications/Law2010DoubleSocialLifeofMethod5.pdf [Accessed January 10, 2012].

Leydesdorff, L. & Welbers K. (2011) 'The semantic mapping of words and co-words in contexts', *Journal of Infometrics*, vol. 5, no. 3, pp. 469-475

Lezaun, J. (2007) 'A Market of Opinions: The Political Epistemology of Focus Groups', *The Sociological Review*, vol. 55, pp. 130-151.

Lin, Y. (2012) 'Transdisciplinarity and Digital Humanities: Lessons Learned from Developing Text-Mining Tools for Textual Analysis', In: *Understanding Digital Humanities*, ed. D. Berry, Palgrave, Basingstoke.

Loukides, B.M. (2010) 'What is Data Science?', *O'Reilly Radar Report,* 2 June. Available at: http://radar.oreilly.com/2010/06/what-is-data-science.html [accessed at January 7, 2012].

Lynch, M. (1991) 'Method: measurement – ordinary and scientific measurement as ethnomethodological phenomena', in *Ethnomethodology and the Human Sciences*, ed. G. Button, Cambridge University Press, Cambridge, pp. 77- 108.

Manovich, L. (2011) 'Trending: The Promises and the Challenges of Big Social Data', *Unpublished Ms*, pp.1-17. Available at: http://www.manovich.net/DOCS/Manovich_trending_paper.pdf.

Manovich, L., Douglass, J., Zepel, T (2012) 'How to Compare One Million Images?', In *Understanding Digital Humanities*, ed. D. Berry, Palgrave Macmillan, London.

Marres, N. (Forthcoming) 'Re-distributing methods: Digital research as participatory research', *The Sociological Review*. Available at: http://eprints.gold.ac.uk/6086/1/Marres_re-distributing_methods.pdf.

Marres, N. (2012) 'The environmental teapot and other loaded household objects: Re-connecting the politics of technology, issues and things', In *Objects and Materials: A Routledge Companion*, eds P. Harvey, E. Casella, G. Evans, H. Knox, C. McLean, E. Silva, N. Thoburn & K. Woodward. Routledge, London.

Marres, N. & Rogers, R. (2005) 'Recipe for Tracing the Fate of Issues and their Publics on the Web', In *Making Things Public: Atmospheres of Democracy*, eds B. Latour & P. Weibel, MIT Press Cambridge, MA.

Moens, M.F. (2006) *Information Extraction: Algorithms and Prospects in a Retrieval Context*, The Information Retrieval Series 21, Springer, New York.

Mohr, G., Stack, M., Ranitovic, I., Avery, D. & Kimpton, M. (2004) 'An Introduction to Heritrix: An open source archival quality Web crawler', *4th International Web Archiving Workshop (IWAW04)*.

Newman, M.E.J., Barabási, A.L. & Watts, D.J. (2007) *The structure and dynamics of networks*, Princeton University Press, Princeton.

Niederer, S. & Van Dijck, J. (2010) 'Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system', *New Media & Society*, vol. 12, no. 8, pp. 1368-1387.

Page, L., Brin, S, Motwani, R & Winograd, T. (1998), 'The PageRank citation ranking: Bringing order to the Web', *Stanford Digital Libraries Working Paper*.

Quine, W.V.O. (1951) 'Two Dogmas of Empiricism', *The Philosophical Review*, vol. 60, pp.20-43.

Rheinberger, H.J. (1997) *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press, Stanford.

Rieder, B. & Rohle. T. (2012) 'Digital Methods: Five Challenges', In *Understanding Digital Humanities*, ed. D. Berry. Palgrave, Basingstoke.

Rogers, R. (Forthcoming) *Digital Methods*, MIT Press Cambridge, MA.

Rogers, R. (2005) 'New media effects: Do formats organize networks?', *Complexity*, vol. 10, no. 5, pp.22-34. Available at: http://doi.wiley.com/10.1002/cplx.20083 [Accessed January 7, 2012].

Rogers, R. (2002) 'The Issue Crawler: The Makings of Live Social Science on the Web' *EASST Review*, vol. 21, no. 3/4, pp. 8-11.

Sarawagi, S. (2007) 'Information Extraction', *Foundations and Trends in Databases*, vol. 1, no. 3, pp.261-377.

Savage, M. & Burrows, R. (2009) 'Some Further Reflections on the Coming Crisis of Empirical Sociology' *Sociology*, vol. 43, no. 4, pp.762-772. Available at: http://soc.sagepub.com/cgi/doi/10.1177/0038038509105420 [Accessed August 15, 2011].

Savage, M. & Burrows, R. (2007) 'The Coming Crisis of Empirical Sociology', *Sociology*, vol. 41, no. 5, pp.885-899. Available at: http://soc.sagepub.com/cgi/doi/10.1177/0038038507080443 [Accessed July 20, 2011].

Shah, C. & File, C. (2011) 'InfoExtractor – A Tool for Social Media Data Mining', *JITP 2011: The Future of Computational Social Science*. Available at: http://scholarworks.umass.edu/jitpc2011/7/ [Accessed January 8, 2012].

Thelwall, M. (2001) 'A Web crawler design for data mining', *Journal of Information Science*, vol. 27, no. 5, pp.319-325. Available at: http://jis.sagepub.com/cgi/doi/10.1177/016555150102700503 [Accessed September 3, 2011].

Uprichard, E. (forthcoming)'Being stuck in (live) time: The sticky sociological imagination', *The Sociological Review*, Special Issue on Live Methods, eds L. Back & N. Puwar.

Uprichard, E., Burrows, R. & Byrne, D. (2008) 'SPSS as an "inscription device": From causality to description?', *The Sociological Review*, vol. 56, no. 4.

Uprichard, E. (2012) 'Dirty Data: Longitudinal classification systems', *The Sociological Review,* Special issue on Measure and value, eds L. Adkins & C. Lury.

Venturini, T. (2010) 'Building on Faults: How to represent controversies with digital methods', *Public Understanding of Science*, December.

Webb, E.J., Campbell, D.T., Schwartz, R.D. & Sechrest, L.: (1966) *Unobtrusive measures: Nonreactive research in the social sciences*, Rand McNally College Publishing Company, Chicago.

Webber, R. (2009) 'Response to "The Coming Crisis of Empirical Sociology": An Outline of the Research Potential of Administrative and Transactional Data', *Sociology*, vol. 43, no. 1, pp.169-178. Available at: http://soc.sagepub.com/cgi/doi/10.1177/0038038508099104 [Accessed August 15, 2011].

Weltevrede, E. (Forthcoming) 'Repurposing Google for national Web research', in *National Web studies: Digital methods to locate, demarcate and diagnose the condition of the national from a Web perspective,* PHD dissertation, University of Amsterdam.

Weltevrede, E. (2011) 'Digital methods to study digital natives with a cause', In *Digital Alternatives with a cause?*, eds N. Shah & F. Jansen, Bangalore & The Hague, pp. 10-23.

Woolgar, S. (1988) *Knowledge and Reflexivity: New Frontiers in the Sociology of Knowledge,* Sage, London

Whatmore, S.J. (2009) 'Mapping knowledge controversies: Science, democracy and the redistribution of expertise', *Progress in Human Geography*, vol. 33, no. 5, p.587.