Journal of Cultural Economy
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rjce20

SCRAPING THE SOCIAL?

Noortje Marres & Esther Weltevrede

Published online: 22 Feb 2013.

To cite this article: Noortje Marres & Esther Weltevrede (2013) SCRAPING THE SOCIAL?, Journal of Cultural Economy, 6:3, 313-335, DOI: 10.1080/17530350.2013.772070

To link to this article: http://dx.doi.org/10.1080/17530350.2013.772070

SCRAPING THE SOCIAL?
Issues in live social research

Noortje Marres and **Esther Weltevrede**

(Received 4 Apr 2012; Accepted 13 Dec 2012)

This paper investigates the device of scraping, a technique for the automated capture of online data, and its application in social research. We ask how this 'medium-specific' technique for data collection may be rendered analytically productive for social research. We argue that, as a technique that is currently being imported *into social research, scraping has the capacity to re-structure research in at least two ways. Firstly, as a technique that is not native to social research, scraping risks introducing 'alien' analytic assumptions such as a pre-occupation with freshness. Secondly, to scrape is to risk importing into our inquiry categories that are prevalent in the social practices and devices enabled by online media: scraping makes available* already formatted *data for social research. Scraped data, and online social data more generally, tend to come with analytics already built in. The pre-ordered nature of captured online data is often approached as a 'problem', but we propose it may be turned into a virtue, insofar as data formats that have currency in the practices 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 'live' social research, in which the formats and life cycles of online data may lend structure to the analytic objects and findings of social research. We demonstrate this point in an exercise of online issue profiling, and more particularly, by relying on Twitter and Google to track the issues of 'austerity' and 'crisis' over time. 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; live sociology; digital methods; automated information extraction; digital social research; science and technology studies; information formats

1. Introduction

Scraping, to state this quite formally, is a prominent technique for the automated collection of online data. It is one of the most distinctive practices associated with current forms of digital social research, those that are marked by the rise of the Internet and the new ubiquity of digital data in social life. Scrapers, to say it more informally, are bits of software code that makes it possible to automatically download data from the Web, and to capture some of the large quantities of data about social life that are available on online platforms like Google, Twitter and Wikipedia. Scraping is widely seen as offering new opportunities for digital social research: it promises to enable the development of new ways of collecting, analysing, and visualising social data. These opportunities have been the subject of a fair amount of hype in recent years: they have been advertised in various

programmatic pronouncements on the future of digital social research under labels like 'the computational turn' in social research and 'big data'. As our entry point into this area, however, we wish to focus in this article on the relatively mundane practice and device of 'scraping' itself. As a technique of online data extraction, scraping seems of special interest to us because it is an important part of what makes digital social research *practically* possible.

By way of context, it should be noted that, besides being a prominent practice, scraping is today also a notable news item. Scraping is granted special importance in stories in the media about the 'revolution' in social research enabled by the Internet. Newspapers from the *New York Times* to the *Wall Street Journal* have recently run articles on scraping, making dramatic pronouncements about its social, economic, and epistemic implications. As the *New York Times* reported,[1] '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' (Markoff 2011). 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'.

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 techniques 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 (Back 2010; Savage *et al.* 2010). It is to adopt a more myopic focus on the concrete and everyday techniques and practices of online digital data capture. Such an approach allows us to take up central ideas from the sociology of science and technology and apply these to, and in, digital social research.

An investigation of scraping as a technique of social research allows us to renew a commitment long held dear by sociologists of science and technology, namely the insistence that science is best understood as a materially specific practice (Knorr-Cetina 1981; Latour & Woolgar 1979; Haraway 1997).[2] When considering a research technique like scraping, we soon find ourselves describing the particular settings in which digital social research is done. Scraping also allows us to approach digital social research as an on-going process, rather than as a finished product (Lury 2012): it opens up a perspective on 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 (Latour 1987). But perhaps most interesting of all, we find, is that these sociological ideas may not only be *applied to* scraping, they may also be *deployed in* scraping-enabled social research.

In what follows, we will argue that digital social research offers ways of renewing the commitment to research-as-process, as scraping may inform the development of 'live' forms of social research, a term we take from Lury (2012) and Back and Puwar (2012). Crucial in this respect is that scraping disturbs the distinction between the 'inside' and the 'outside' of social research. The development of scraping as a data collection technique has been largely exogenous to academic social science: its recent popularity is closely associated with the rise of the so-called 'real-time Web,' which is, as we will explain, an industry term. Scraping is arguably enabling a distinctive approach to knowledge-making across social life, one that is pre-occupied with monitoring 'what is happening, right now,' to use the slogan of the micro-blogging platform Twitter, a favourite target of scrapers. This circumstance defines both the key challenge and opportunity of scraping for social research: if scraping is already deployed in the 'real-time Web', what does social research hope to contribute by taking up scraping? To what additional or alternative purposes might scrapers be put in/by social research?

In the second half of the article, we will present some of the distinctive applications of scraping developed in recent social research. The most interesting of these applications, in our view, do not ignore the fact that scraping, as a data collection technique, is not native to social research. There is something interesting about the ways in which scraping introduces 'alien' assumptions into social research. Scraping does not just insert a strange technique into social research practices, it involves the importation of categories into social research that are strictly speaking external to it: scraping makes available *already formatted* data for social research. We will argue that this makes possible a distinctive approach to social research, one which approaches the formatting of online data as a source of social insight, and which we call 'live' social research.

We will develop this argument through a discussion of diverse empirical materials. In keeping with our understanding of scraping as a multi-faceted device, we assemble our account of the technique from a variety of sources: from scraping manuals to popular scraping-based research, to technical scientific articles on the subject. Once we have established the features of scraping as a research technique and practice, we will present a pilot study, one in which we scraped Google and Twitter for particular keywords, 'austerity' and 'crisis'. This pilot study should give, if not proof, then at least a taste of the distinctive form of 'live' social research that scraping enables. Provided that we learn how to take advantage of the data formats that scrapers help make accessible for analysis, scrapers offer a powerful instrument for social research, one that extends our grasp precisely *beyond* what is happening 'right now'. But first we need to talk about the scrapers themselves.

2. What is Scraping?

It seems good to begin our investigation with an actual example: in Figure 1 you can see a scraper at work. This screenshot provides a live view of scraping: it shows the moment when the scraper script is running and busy extracting data from the World Wide Web. The particular scraper in the screenshot is pulling information off the online encyclopaedias Wikipedia. It is querying a specific page, one called 'List of Occupy movement protest locations', to extract from it a list of cities and towns where 'Occupy' protests were being held at that time. This sounds relatively straightforward, and indeed it is, insofar as this scraper is scraping only this one page, of which the title already describes

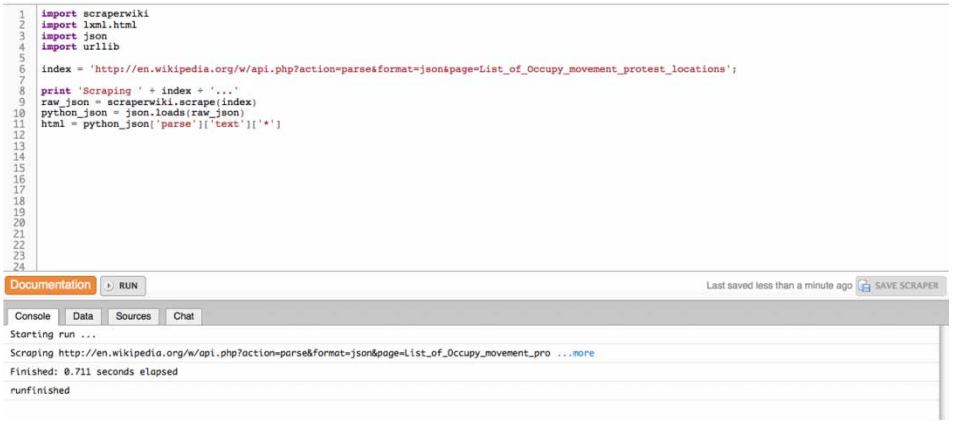

FIGURE 1
Live view of the ScraperWiki

the type of data we are scraping for: a list of cities and towns. However, the example also raises some less straightforward questions. Scraping is usually described as a technique for data collection, but isn't this scraper also *analysing* data? After all, this script is not extracting data *indiscriminately* from the Web, but only information on a particular page that fits a particular category: 'Occupy movement protest locations'.

Precisely this capacity of scraping to extract 'structured information' from sources is highlighted in the formal definition of scraping, in the technical literature on 'informational retrieval'. While Web scraping is specifically targeting online data, automatic data capture has a much broader application, and a much longer history, going back to at least the 1970s. Information extraction has been generally defined as 'the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources' (Sarawagi 2007, p. 261). Arguably, it is precisely insofar as information retrieval relies on *information structures* to extract data that it has acquired special saliency in the current context of the digitisation of social life.[3] In this context, the issue of information overload and the availability of overwhelming quantities of data are widely regarded as a central challenge and opportunity (Moens 2006). Information extraction promises a medium-specific solution to this problematic, in that it offers a way to extract – quite literally – relevant information from the data deluges enabled by digital and networked media. On the Web, information extraction addresses the problem of relevance *by deriving ordered data out of the heterogeneously ordered expanse of digital information environments*. It offers a solution as it provides a way of extracting specific fields or data elements from pages on the Web and other Internet sources, turning online data into usable, well-ordered data sets.

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. One can scrape Web pages for images, or location data (town, region, country) or for 'keyword in context' data (bicycle: 'Find cheap *bicycles* for sale in London'), which can then serve as a data set for research.[4] Insofar as scraping relies on information structures, the technique is easiest to use if the targeted content *itself* has structure, for instance, if it takes the form of a table, or is formatted as 'tweets'. In another

sense, however, scraping treats the Web blindly, as a limitless expanse from which only specific elements need to be brought in. 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.

Scraping, however, is not only a technique but equally involves a particular way of dealing with information and knowledge: it is also an *analytic practice*. A lot of 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 the micro-blogging platform Twitter. Scraping presupposes *a wider socio-technical infrastructure*, and especially important in this regard is 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' (Watters 2011). To put this differently, the rising 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).[5] In large part because of its very freshness, this data is in need of constant disclosure and analysis. Scraping is the way to capture these fresh online data and there is a variety of tools and projects available online offering 'real-time analysis' which have been made using scrapers. Figure 2 provides an example of the use of scraping in 'live' news reporting: it is a visualisation 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 mention on the online

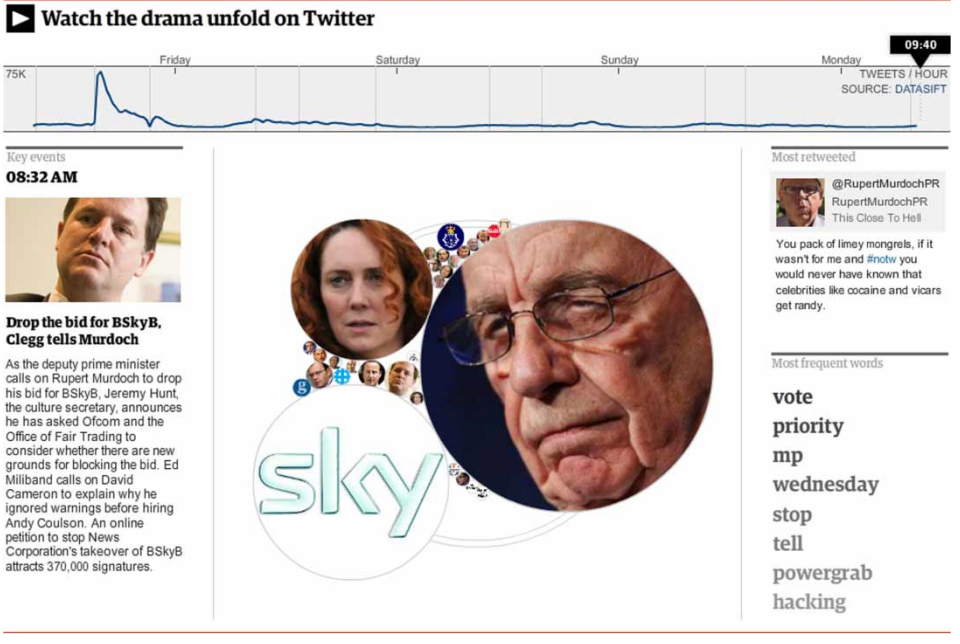

FIGURE 2
How *The Guardian* tracked the *News of the World* scandal on Twitter

chatter channel, thus providing an account of the scandal 'as-it-happened' (Richards *et al.* 2011).

Another example of the type of analytic practice enabled by scraping is ScraperWiki: this is a platform for developing and sharing scrapers from which the example in Figure 1 was taken. ScraperWiki turns scraping into a service, as the platform 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. On its site, ScraperWiki explicitly frames its services as including data provision for media, organisations and government. ScraperWiki also responds to a common problem with scrapers: their instability.[6] 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 or the changes in layout and design of the pages and sites to be scraped. 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 'plastic methods' (Martina Mertz, quoted in Helmond 2010).

In an ostensible effort to address this situation, ScraperWiki provides a Web-based, generally available platform that makes it more easy to use, develop, archive and manage scrapers. This also suggests it would be a mistake to approach scrapers as if they were stable, stand-alone machines: scrapers come in and fall out of use; they work, and then they no longer work. Insofar as scraping involves the on-going collaborative practice of sharing, editing and copying of scraper code, it also involves a distinctive set of everyday social practices, as one of us discovered while learning how to scrape with the ScraperWiki in an enjoyable lab session with MA students at Goldsmiths, University of London. This session also helpfully presented us with a scraping 'recipe,' which details the relatively simple steps students were to follow in order to extract a particular type of data from the online encyclopaedia Wikipedia (as described in the example at the beginning of this section) (see Figure 3).

As this recipe suggests, scraping involves a series of steps in which formatted data is derived from an informational mess. To scrape is to build a chain from the heterogeneously formed mass of online data to formatted information, and along this chain Web 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'). After 'running the scraper', a series of further steps follow, in which the data is cleaned in successive operations, removing redundant html code and other irrelevant bits, until only the targeted data remains. This is how scraping structures knowledge-making as a distillation process. As such, scraping calls to mind the epistemic category of the 'plasma' usefully defined by Emmanuel Didier (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 messy, partly opaque streams, and then relying on these formats to extract analytically useful data for research (see on this point also Latour 2005).

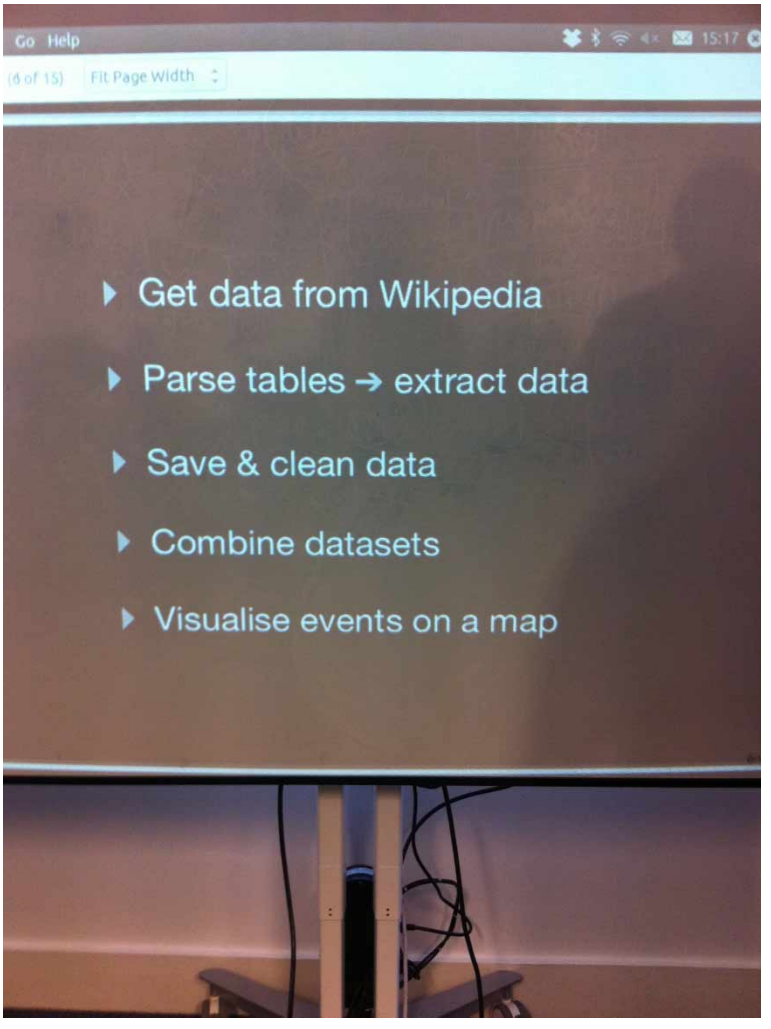

FIGURE 3
Learning how to scrape with ScraperWiki in the CAST Lab, Goldsmiths (December 2011)

Scraping, then, is a multi-faceted phenomenon. It offers a technique for automatically capturing online data, but it is also an important component of a broader analytic practice, one that is marked by the rise of the 'real-time Web,' and the dynamic kinds of information services it enables. Scraping is also a social practice, one in which teachers and students share recipes, and programmers and researchers exchange scraper data and applications. Finally, scraping seems to imply a distinctive approach to knowledge-making. It arguably comes with *an epistemology built in*: scraping formats the process of data collection and analysis as an operation of extraction, and organises knowledge-making as a distillation process. One could say that the combination of these different features is what makes scraping not simply a technology, but a 'socio-technical device' as sociological parlance has it. The various features of scraping are of sociological interest in and of themselves, but we feel they are especially important to understanding

how scrapers can be *deployed* in social research – empirically, analytically and normatively speaking.

3. Scraping as a Device of Social Research

It is also clear that scraping is not native to social scientific research. To use scrapers in social research is to import a technique from the worlds of information science and digital services, one that must be adapted if it is to suit the purposes of social research (just as scrapers themselves are devices for re-purposing online data). The precarity of scrapers as techniques of social research is underlined by 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, as their name says, commonly use scrapers.[7] Additionally, in some online discourses scraping is most closely associated with illegal practices, because scraping may be against the terms of use. Taking up scrapers, social researchers find themselves in the position of having to adapt to their needs a technique that already serves a broad range of purposes in social life, many of which aren't exactly reputable.

Many of the defining methods and techniques of social research are assumed to have been born and bred in the social sciences, even though this pedigree is easily overstated. A relevant example here are the techniques associated with the previous 'computational turn' in social science: with the rise of the software programme SPSS in the 1970s and the statistical analysis of survey data that it enabled (Uprichard *et al.* 2008). The SPSS package is today used in a variety of fields, from marketing to healthcare governance, but the technique is generally understood to be of sound social scientific pedigree, having been developed at the University of Chicago, in their National Opinion Research Center. It is certainly possible to question the 'purity' of this social scientific origin – one of the developers of SPSS headed the Computation Centre at Chicago and SPSS was packaged as a commercial product almost from the very start – but the very fact that this is something to 'bear in mind' tells us that the heterogeneous origins of social research techniques are less than obvious. By contrast, in the case of scraping it is all too evident that the technique has been developed elsewhere. Social research applications of scraping re-purpose widely used techniques of data collection and processing, and *noticeably* so.

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 & File 2011). While the methodological purpose of this scraper is solidly social scientific, namely discourse analysis, the way it works is not unlike the techniques developed under the rubric of 'Facebook analytics' by software developers, many of whose interests do *not* qualify as social scientific (but rather resemble those associated with the ScraperWiki discussed above). In our view, the more interesting uses of scraping in social research do *not* disavow these continuities between scraping as a social and computational practice and its applications in social research, but to the contrary, seek to take advantage of them, analytically and empirically. Some social research applications of scraping are very explicit about the fact that they are re-purposing popular technology. A notable example here is Google Scraper, a tool for online textual analysis developed by the Digital Methods Initiative in Amsterdam.[8] This scraper pulls

information from the search engine Google, extracting parts of the query return pages from it.

Google Scraper has been explicitly presented as a way of repurposing Google as a 'research tool' (Rogers 2009; Weltevrede forthcoming).[9] This scraper provides a way to collect and analyse Google return pages over time, as it allows researchers to automatically query Google for specific keywords. As such, Google Scraper offers a means to study Google itself, in particular the ways in which it ranks Web sources on its return pages, which are so influential in structuring traffic and 'attention' online. However, in another application of Google Scraper, the tool can also be used by social and cultural researchers to perform textual analysis on the Web. Google Scraper makes it possible to detect the presence of specific words on particular sets of Web pages selected by the researcher, and the scraper then relies on the search engine's index and search operators as a vehicle for conducting a basic form of social research: word frequency analysis. Figure 4 provides a snapshot of an initial output of a Google scrape, which shows the presence and frequency of occurrence of issue terms on a particular Web page: in this case, that of an advocacy organisation campaigning against a European Intellectual Property Law, called ACTA, in July 2012 (see Figure 4).[10] In this second application, Google Scraper explicitly re-purposes available technologies and data for social research. As such, it also brings into view some of the wider methodological issues that scraping poses for social research.

4. Scraping and the Re-distribution of Social Research

Google Scraper highlights a particular conundrum that has long concerned social researchers and theorists: where does social research derive its analytic categories from? From social theory or from the social practices under study? In the past, sociologists have taken a strong position on this issue, claiming that social researchers should not assume that the social world fits their own categories, but should 'follow the actors', and carve up the world as they do (Becker 2007; Latour 2005). Arguably, digital social research displaces this debate about the provenance of the analytic categories of social research onto the plane of devices: the question here becomes how the categories and formats *implicit in digital technologies* structure our social data and analysis. The scrapers discussed above derive at least some of their formats from the online devices and content queried by them:

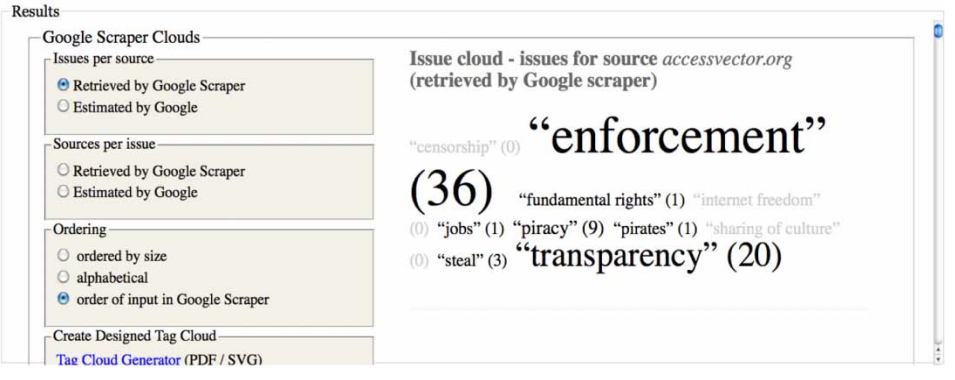

FIGURE 4
A Google scraper result: A set of anti-ACTA web pages queried for ACTA issues

the table with towns and locations on Wikipedia, the comments on Facebook, or the ranked page lists of Google. Scraping-enabled social research tends to adopt analytic categories that have acquired saliency both in the technologies it deploys *and* the practices these technologies enable.

The question is how far we should go in taking online devices and platforms into account as notable components in digital social research? Can we simply understand these devices and platforms as part of our 'methodology', or should we insist that they are part of the 'object' of our analysis? Scrapers raise the question to what extent we are studying the platform being scraped, such as Google or Wikipedia, or the forms of social life supposedly enabled by these platforms (Weltevrede forthcoming). Scrapers, that is, force the question of how we establish the difference between *researching the medium* and *researching the social.* Scraping suggests that this distinction may be a lot less stable or robust than we are inclined to think. And this in turn, opens up a further question: why would we as social researchers ever accept this muddying of the waters? Why would we allow our definition of social life, as something distinct from or exceeding media accounts of it, be unsettled in this way? In this respect, it is important to recognise that scraping has a number of distinctive affordances for social research.

First and foremost, scraping solves a problem that social research shares with many other digital practices: it offers 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. This is why scraping has been said to do no less than to unlock the 'sociological potential' of the Web: scraping promises to make available for social research the very large quantities of user-generated data that currently are being amassed through online platforms. Crucially, however, it should be noted that the 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, most notably by offering APIs, so-called application programming interfaces. 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. Social research that relies on APIs risks rendering itself platform-dependent, and in effect accepts the blackboxing of its data collection methods (boyd & Crawford 2011).

Scraping also holds a second attraction for social research: 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 'incomplete', 'messy', and 'tainted' data (Savage & Burrows 2007; Uprichard 2012). Such statements make sense if we compare online data with the data sets that many social researchers are most used to working with, such as social survey data, and interview transcriptions. However, from the standpoint of scraping, these characterisations of online data are decidedly odd. As we noted above, scraping offers a way to extract structured information from heterogeneously formatted data online. From this standpoint, it seems a mistake to say that 'online data' is inherently this or that. Scraping highlights that the quality of online data is in part a *processual accomplishment*: it partly depends on the operations that devices and researchers perform on Web data, how good or clean these may become. In this respect, scraping also re-opens the debate about the techniques that are used in social research for protecting and/or enhancing the quality of social data (Webber 2009; see also Gros 2012).

Scraping, we think, invites us to re-frame or re-locate the sociological concern with the quality of online data. The widespread application of scraping across social life reminds us that the problem of 'dirty data' is not at all exclusive to social and cultural research. Many professions are 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 amenable to research (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 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 a vast collection of 'completely uncontrolled heterogeneous documents' both in terms of internal variations and external meta information (Page *et al.* 1998). It was in this context that scrapers (and crawlers) emerged as devices capable of bringing order to – or, rather, extracting order from – the Web, as they make it possible to collect and re-structure large quantities of heterogeneous sources to be queried.[11]

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 both its devices and its research challenges with a host of other actors, technologies and agencies that constitute the digital networked environment. Scraping is indicative of a wider 'redistribution' of social research that is enabled by digital technology (Marres 2012): many entities and issues that are conventionally located outside social research come to contribute actively to the performance of social research. To adopt this standpoint has implications for how we understand the relations between the inside and outside of social research. First, 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, scraping invites us to approach online 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.[12] Entities that are in some respects alien to the context of academic social research may come to play a noticeable role in its organisation, with implications for analysis.

Scraping also suggests a particular take on the relation between the *objects* and *methods* of digital social research, and this is where the sociological concept of the device proves especially useful. In the philosophy and sociology of science and technology, this concept has been used precisely to highlight the relative *fluidity* of the distinction between object and method (Duhem 1954; Latour & Woolgar 1979; Rheinberger 1997). This insight has also been applied to social research. Once we approach social research from the standpoint of its apparatus, the distinction between the techniques, methods, and objects of research becomes hard to sustain. If we consider methods like the 'social survey' or the 'focus group' as a located practice, 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 fitting instantiation of this argument (as well as of other STS arguments, see Rieder & Rohle 2012). In the case of scraping, too, it seems impossible to make straightforward distinctions between the instruments, methods and objects of

digital social research. Google Scraper, for example, outputs its data as a structured text file (with categories) and in the form of a resonance or word frequency analysis. In this respect, it seems technically wrong to call this scraper a device for data *extraction*: it also performs *analysis* and contributes towards the public presentation of results (in the form of so-called tag clouds) (see Figure 4). More generally speaking, it seems strange to ask whether terms identified by scrapers exist 'independently' from the devices deployed. However, scraping does not only offer a useful demonstration of the entanglement of method, object and technique in online research. It also provides opportunities to *deploy* them in research.

In some respects, it should come as no surprise that these ideas about devices taken from the social studies of science and technology are so easily applicable to scraping. In developing these ideas, STS scholars applied technical metaphors of information processing to the abstract phenomena investigated in the philosophy and sociology of science (proposing, for instance, the aforementioned idea of reference as a chain). Little wonder that these ideas provide a useful language for making sense of the use of information processing techniques in social science! However, here we want to argue that we can derive some useful guidelines from this approach, as to how to *deploy* informational devices in social research. We want to propose that it is *precisely* insofar as 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, scrapers tend to derive their analytic categories from the online platforms queried by them. They use the data formats that are implicit in social data in order to structure these data. This state of affairs, it seems to us, may be *deliberately* deployed in social research for analytic purposes. This can also help to clarify what social research may contribute to scraping-enabled research in its broad definition.

5. Live Social Research: Meta Data as the New Social Data?

To be sure, the issue of 'pre-ordered data' has often been treated as posing epistemic, normative and political challenges for social research (Bowker & Star 1999). However, 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 digital social data has also been apprehended critically by some sociologists, who have flagged difficulties in terms of the limited life span and deterioration of 'live' data sets (Uprichard 2012). However, scraping offers possibilities for turning this analytic vice into a virtue. The fact that data clearly have a life cycle online may be deployed in social research, in a context in which such 'life signals' can be analysed with the aid of scraping techniques.

An example of this type of approach can be found in the study Historical Controversies Now, which compares queries for recent and less recent historical controversies across different platforms – from Twitter to Google Scholar (see Figure 5; Dagdelen *et al.* 2010). The study exposes the different ways in which online platforms structure issues temporally: Twitter organises tweets by freshness (i.e. date stamps), thereby favouring current and contemporary reworkings of past events. By contrast, Google Scholar foregrounds well-cited accounts of these same historic events, something which significantly expands the temporal frame, as citations take time to accrue (See Figure 5). The study demonstrates how online platforms format issues using meta-data

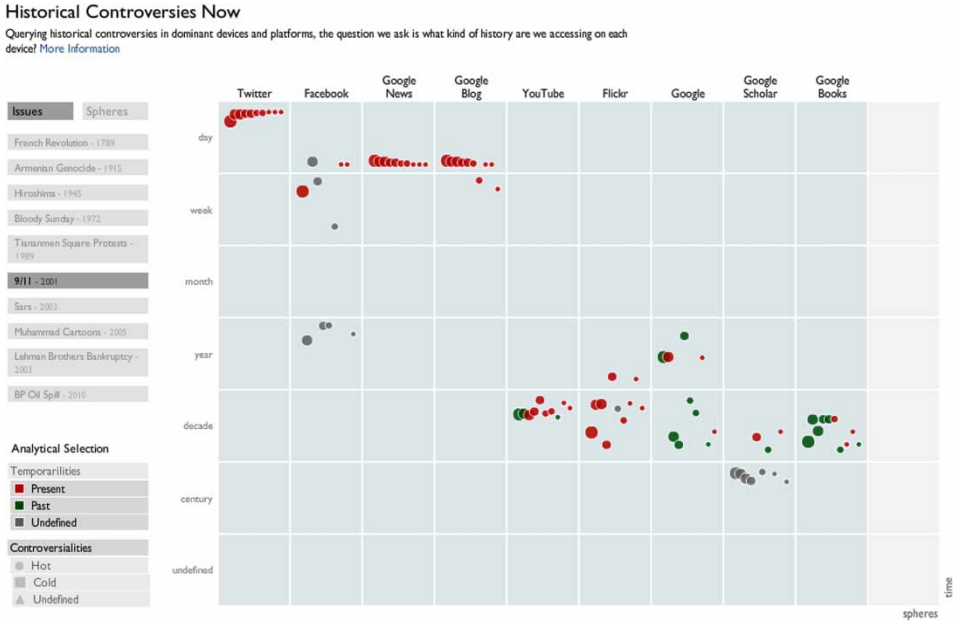

FIGURE 5
Historical controversies now, digital methods initiative (2010)

(date-stamps and citations). However, it also turns the life cycle of data itself into an object and vehicle of analysis: it identifies the temporal envelopes of issues across platforms by plotting the publication dates of sources associated with engine returns for each of these platforms.

We are suggesting, then, that the dynamism of online data, far from being only a problem (deterioration, incompleteness), may itself be deployed for analytical purposes in online research. Sociological concerns about the dynamism of social data out there may be much less transferrable to online data than some sociologists seem to have instinctively assumed. The problem of dynamic data online does not necessarily have to be 'kaltgestelt' – as the German language puts it appropriately – by freezing a data set in time, so as to render it stable (Latour 2005). In the remainder of this paper, we want to substantiate this proposal by outlining a distinctive style of digital social research: live social research.

Liveness is the term proposed by Lury (2012) and Back and Puwar (2012) to investigate the transformation of the spaces and times of social research in the context of digital culture. For them, the term captures the need for sociology to become responsive to contemporary changes in the spatial and temporal ordering of social phenomena: the increasing valorisation of instantaneity and liveness, the drive towards the condensation of the past, present and future in digital networked media, and the conjuring up of an 'eternal now' in this context. Here we would like to suggest that 'live research' can not only be taken to refer to changing relations *between* social research and its social context: it may also indicate an *internal* re-ordering of empirical social research in the digital context. The digitisation of social life, and the method of scraping more in particular, signal a re-ordering not just of the times and spaces of social research, but also *of the empirical cycle* itself.

As discussed, the collection and analysis of data are clearly and explicitly entangled in the process of scraping.[13] Scrapers do not just capture but also 'parse' the data culled from the Web – scrapers are analytical machines. Accordingly, we could say that in scraping-enabled forms of social research, analysis precedes data collection rather than succeeding it. And there are at least two different ways in which social researchers may deal with this circumstance. They may focus on the data, and try to 'clean' or strip the content of its formatting before analysis. Or they may try to take advantage of the analytic features of online data, and treat the formats as 'metadata' that may be operationalised in social analysis. To take seriously the latter possibility is to explicitly reject an aforementioned assumption about digital data in sociology: the suggestion that digital online data sets are an informational mess. As we have seen, from the standpoint of scrapers the opposite is 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. We propose to define as 'live,' social research that seeks to derive its analytic capacities from the device-specific 'pre-formatting' that is distinctive of dynamic online data.[14]

To adopt such an approach is to take a strong position in the debate flagged above, about where social research should derive its analytic categories. Live research values positively the fact that online information extraction techniques pre-format data for social and cultural research and seek to render these formats productive for research (see also Niederer & Van Dijck 2010). This applies to information formats, such as the tweet or the comment, but also to more complex deployments of digital markers. As we saw above, search engines format heterogeneous data sets by collecting and processing data through markers such as hyperlinks and anchor texts, and online platforms pre-format social transactions in software and interface design, such as the re-tweet or like. In both cases the online devices already order, or more precisely, pre-format, the data for social research. Rather than treating these formatting effects as something that contaminates our data with negative bias, we propose that social research may partly derive its analytic capacities from these effects. Live social research asks how data formats that are already implicit in the data may be analytically deployed to structure social research and generate 'findings'.

To take up this project, finally, is to break explicitly with a convention built into classic social research devices: in the statistical analysis of surveys, or the marking up of interview data – and associated tools like SPSS and Nvivo – the analytic categories are predefined by the research design, and the data is made to fit these 'indigenous' categories in the execution of the research. In scraping-enabled research, by contrast, the formatting of data by digital devices 'out there' becomes available as a source of relevant categorisations. It takes up an insight of constructivist methodologies developed in the sociology of science, like citation analysis which propose that formal operators with currency in practices under study, like citations, may be a source of sociological insight (Marres 2012 discusses this more extensively). The focus on the empirical deployment of *device-specific* categories gives yet another twist on the sociological debate about the status of its categories. The aim is not only to determine which formats and categories are analytically most productive for social research, i.e. to determine relevant ways of the carving of the world. It is also a question of *operationalising* devices with currency in social life 'out there' for purposes of social research.

Live social research, then, affirms the re-distribution of social research that is enabled by digital media, granting a prominent role to the data formats and categories that

structure information out there, in the social world. Here, 'extraneous' features of data, like a date-stamp or hyperlinks received, provide key indicators. These data attributes could be called metadata, even if they are conceptually distinct from technical metadata features, such as name, size, data type. Metadata are of course data themselves, but they are also descriptive of, and structure, other data. We call this type of research 'live' social research, as it seeks (1) to 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 date-stamps – to track the dynamic composition of social life and its issues online.

6. Pilot Study: From Live Media to the Liveliness of Issues

Live social research, to summarise, is social research that seeks to render analytically productive the formatted, dynamic character of digital networked data. As such, this research practice endorses the dynamism or 'shape-shifting' of online data, turning these into a resource and an object of digital social research. In this respect, one crucial question is how 'live' social research relates to 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 live social research different? We would like to propose 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*.[15] In live social research, the key question is not what topics, sources and actors have the most *currency* at a given moment ('now'). Instead, the crucial question for those researching social dynamics is which entities are the most 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 by outlining a pilot study that further specifies this difference between researching liveness and liveliness.

In order to establish the distinction technically, methodologically and analytically, we decided to scrape two online platforms, Twitter and Google, for a small number of terms, which were likely – we thought – to display both dynamics, liveness and liveliness: austerity and crisis.[16] As part of our scraping exercise, data on the relative prominence (currency) of these key terms were collected as a matter of course, but we were especially interested in establishing the *variability* of these two key words: how active or 'lively' are they? Could we identify significant fluctuations in the vocabulary associated with 'austerity' and 'crisis' over time? To operationalise this question, we relied on co-word analysis, which is a well-established method for analysing text in both sociological research and content analysis (Callon *et al*. 1983; Danowski 2009), and which has traditionally relied on the formatting of the medium for the structuration of data. Thus, Callon *et al*. (1983) rely on the keywords used to index academic articles in order to detect significant word pairings (co-words) in this literature. In analysing Twitter and Google data, we also derived our units of analysis from information formats that are central to the operation of these platforms: the tweet and the hashtag in the case of Twitter and the 'snippet' for Google, a title and string of keywords that Google returns for each individual page in its query return list and, for both platforms, the date stamp.

One of the major challenges of declining to research liveness, and focusing on liveliness instead, we soon found out, is that we lose a powerful instrument for data reduction. To ask which term is the freshest, or the most popular, we were forcefully reminded, is a splendidly easy way of boiling down vast amounts of data to a few significant words (Emma Uprichard, personal communication.). If we want to determine not the currency of terms, 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? Secondly, in exploring our initial co-word maps, using the network visualisation and analysis software package Gephi, it 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. The very *content* of our pilot study reverberated with the 'language of the wire': with terms that had been occupying the news of the period in question, such as greece, imf, debt, bailout, protest, and the cluster '2012 – davos – economics – gain – pain – misery – brings' (see Figure 6 visualising the 'austerity' issue space on Twitter.)

In an effort to move beyond newsy terms and get to more 'lively' ones, we focused our Twitter analysis on hashtags rather than words, and reduced our time frame to 4 days. Now we did find terms that are more lively both in terms of content and in the sense of fluctuating more strongly, appearing and disappearing on our co-word maps from interval to interval: #wecanbeheros and #screwyouassad in the case of crisis; and #merkelnotmychancellor; #solidaritywithgreece; #banksterganstersin the case of austerity (see Figure 7).[17] The notion that 'social' terms are especially volatile – as compared with terms that figure prominently in the news – was reinforced by our analysis of Google returns for 'crisis' over time. Among the fluctuating terms in this co-word profile are 'planned parenthood' and 'demi moore' and the cluster 'personal; revealing; social; stories' (Slee 2011). This in contrast to the more stable political, economic – and more reliably newsworthy – terms that appear across all or most intervals (debt; euro; syria).

Cruelly, or perhaps justly so, we then found that the more fine-grained co-word variations were much more resistant to visualisation than the broad-stroked dynamics of news terms. As soon as we reduced the quantity of co-word associations in our visualisation, we were likely to lose the more variable and interesting terms from view. In order to mitigate this effect, we will need to further specify our criterion of issue variability, possibly by focusing on terms in the middle, on those words that are neither fully stable nor so volatile as to only appear in burst and which might therefore account for *significant variation* in an issue's composition over time (Marres *et al.*).

7. Conclusion

Our foray into live social research raises as many questions as it answers, perhaps especially that of how to reduce online data in analytically meaningful ways in social research. Insofar as our pilot study throws such questions into relief, however, it indicates some specific avenues for future exploration. One of the main challenges no doubt is that of how to differentiate in more precise ways between the study of media dynamics and social dynamics. Above we proposed a distinction between 'scraping the medium' and 'scraping the social'. The distinction is at the heart of live social research: when we seek to deploy medium-specific features of our data, rather than disavow them, the question inevitably arises whether we are studying social life or rather the media that enable it only

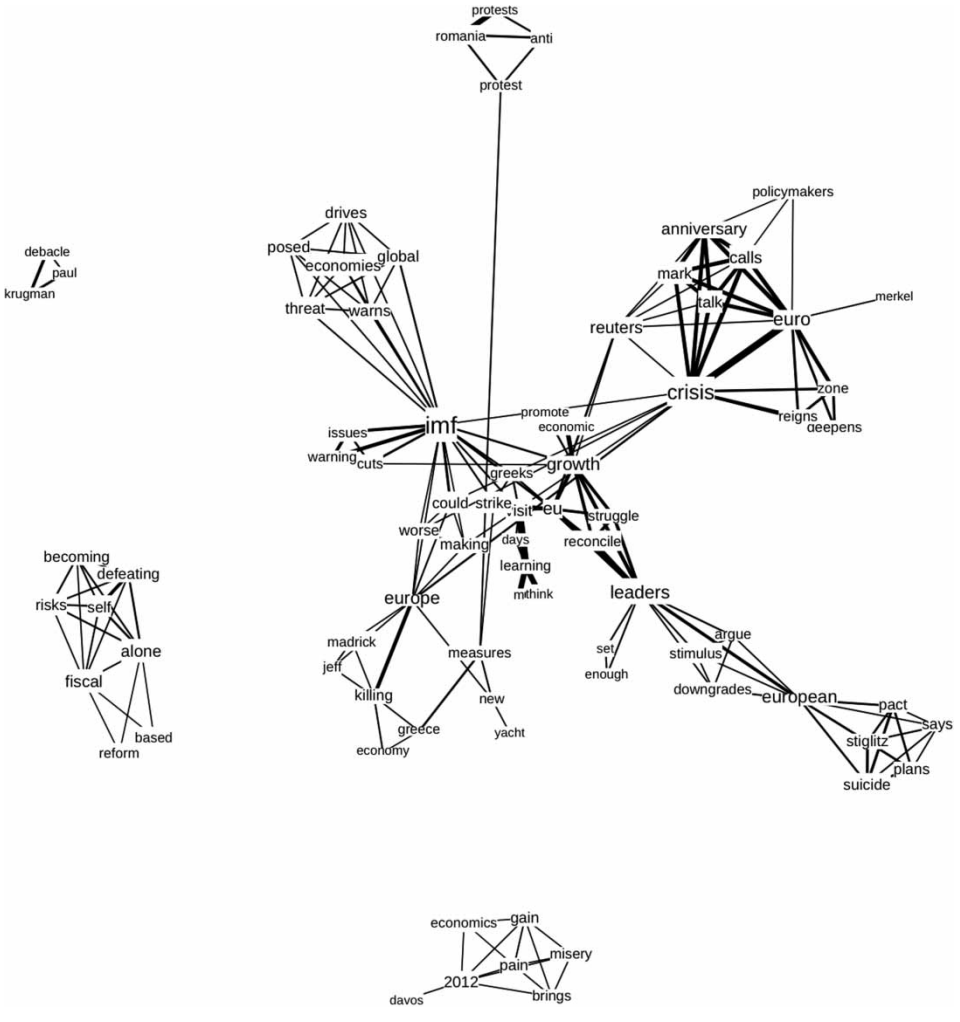

FIGURE 6
Co-words related to austerity derived from the Twitter stream, 1–31 January 2012

in part. But *the difference between 'scraping the medium' and 'scraping the social' is probably best understood as a difference in 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 scraper-enabled research, the boundary between the media apparatus and the social object must be understood as flexible, and to move between 'all apparatus and no object', to 'part apparatus, part object'. Indeed, this distinction may well have to be regarded as a methodological accomplishment. This may be one of the principal merits of live research: as we learn to deploy device-specific attributes in social research, we are forced to keep asking: does our research design mainly help to bring into view the dynamics of media platforms or also those of the social phenomena they enable?

Live social research, finally, also unsettles the distinction between social research and social life. In particular it opens up for questioning the strict separations most of us

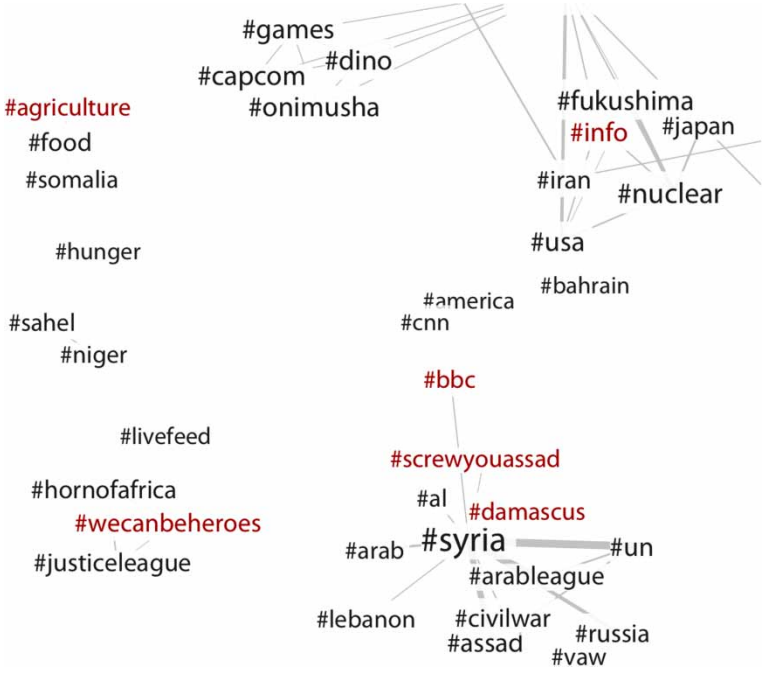

FIGURE 7
Snapshot of co-words related to 'crisis' extracted from Tweets, 24 January 2012

have learnt to make between the epistemological issues that trouble scholars and scientists and the real-world concerns with which social actors are concerned. The methodological and conceptual problems raised by live social research look surprisingly like the issues that trouble digital societies. As we have discussed, one of the big problems of live social research is how to gain full and reliable access to data collected on online platforms. Thus, danah boyd and Kate Crawford (2011) have drawn attention to the relative obscurity of and general disappointment about 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 equally pleasurable ways, highlighted the problem of data-centrism, which is well characterised by Tom Slee in his 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.

While not necessarily solving this problem, live social research entails a change in status of this type of epistemic trouble. In live social research, epistemic issues – platform-dependency, data-centrism, blackboxism – trouble research in rather immanent fashion: they affect it from the inside out and may also become an object of social research. The case study Historical Controversies Now turned platform-dependency into a topic of investigation. And our study of austerity and crisis on Twitter and Google highlighted the ways in which a pre-occupation with liveness is co-produced by data, platforms and research. The epistemic issues that arise in digital social research aren't so different from the issues that trouble digital social life much more generally understood. Our reference to the issues of 'live' social research in the title is then deliberately ambiguous. They refer as

much to the substantive concerns we may examine in online social research (tiananmen square, crisis), as to the methodological trouble that this type of research brings into focus.

ACKNOWLEDGEMENTS

The authors wish to thank the anonymous *JCE* reviewers, Evelyn Ruppert, and participants in the Digital Methods Winter School 2012 for helpful comments on this paper, especially Erik Borra, Carolin Gerlitz and Bernhard Rieder.

NOTES

1. This article refers to scraping, but also to a wider set of techniques used for mining the vast collections of online data, such as through the use of APIs and in-house access to data. Web companies such as Facebook for example have in-house sociologists that have access to the data without using scrapers. Scrapers are thus an 'outsider' technique, used to collect generally accessible data online. Our somewhat artificial focus on scraping seeks to stress the analytic advantage of the pre-ordered nature of online data to the scraper-assisted researcher.
2. 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, these attempts were not so much concerned with the *technological* minutiae of doing social research.
3. Also of relevance here is the semantic Web: the specification of content structures (e.g. title, summary, rating) and annotation structures which would assist in data disclosure and analysis. We however are interested here in *emergent* structures in Web data, rather than efforts to implement them along such systemic lines.
4. There is some debate whether Web crawlers should be included among Web scrapers. Web crawlers 'scrape' links from Web pages and use these links as input to scrape more links. In this paper we will limit our discussion somewhat artificially to scrapers.
5. The real-time Web is closely related to technical developments of the Web (e.g. ajax, feeds, push notifications, APIs), and is seen most notable in Twitter, Facebook's news feed and Google's increasing preference of freshness in search results. see Leggetter (2011).
6. As Helmond (2010) notes, 'Internet methods are incessantly volatile due to the update culture of the Internet itself.'
7. The Wikipedia article on scraping is categorised under 'spamming' and scraping is also discussed in these terms on Google blogs; one of Google's recent algorithm updates, Panda, was designed to combat aforementioned scraper sites. See Wikipedia (2012)
8. Google Scraper is, less transparently, also known as the Lippmannian Device. Available at: http://tools.issuecrawler.net/beta/scrapeGoogle/ (accessed 12 January 2012).
9. See on this point also Borra and Stevenson (2007).
10. This output is a small part of a larger scrape, which used Google to query 59 pages for the list of issue terms related to 'ACTA' presented in Figure 4.
11. Initially Google's algorithm ordered sources purely by popularity measures, built around crawling and scraping hyperlinks and text around links, but increasingly Google is including other signals in their algorithm, such as freshness and personalisation. Google is thus more and more the search engine for 'hot' issues. An important moment in terms

of Google becoming realtime is their Caffeine update, as it makes it possible to offer fresh content almost instantaneously.

12. To be sure, the same can be said of other social research methods. For example, classification practices of surveys or censuses draw from multiple governmental and commercial communities of practice (see for example Bowker & Star 1999). We are insisting here that scraping is *noticeably* an open-endedness practice, and that this has consequences for our understanding and deployment of social research methods online.

13. This relative opacity of the scraping process also raises the issue of transparency in digital social research. A significant amount of online social research is based on non-disclosed data sets, and it has been argued that this opacity should be explicitly challenged: good digital social research should open source the code, or at least provide pseudo code explaining the full recipe of extracting, cleaning and ordering the data. While we are broadly in support of these arguments, we here want to call to mind that the relative opacity of the research apparatus has been a key issue in the sociology and philosophy of science during the 20th century, and as such, represents a problematic we may well need to come to terms with and accommodate in our research practice, rather than bracket it through an appeal to the ideal of transparency. We are in favour of openness on different grounds, namely in order to support the understanding of research as process, and to multiply the opportunities for the empirical articulation of epistemic issues in digital social research.

14. That metadata formats objects of research is well established in other forms of social science research such as surveys or censuses and has been a source of analytic and empirical insight. We are pointing to here to the distinctive forms of device-specific pre-formatting enabled by digital networked media: formats that are deployed 'out there' and 'live' such as date-stamps and hyperlinks. Thanks to Evelyn Ruppert for this clarification.

15. 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).

16. 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.'

17. 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.

REFERENCES

BACK, L. (2010) *Broken devices and new opportunities: re-imagining the tools of qualitative research*. NCRM Working Paper. NCRM. (Unpublished)

BACK, L. & PUWAR, N. (2012) 'Introduction', *Live Methods*, eds L. Back & N. Puwar, Sociological Review Monographs, Wiley-Blackwell, Hoboken, pp. 6–17.

BECKER, H. (2007) *Telling about Society*, Chicago University Press, Chicago.

BERRY, D. (2011) 'Real-time streams', *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, Queenstown.

BORRA, E. & STEVENSON, M. (2007), 'Repurposing the Wikiscanner: comparing Dutch universities' edits on Wikipedia', Masters of Media Blog, 7 October [Online]. Available at: http://

mastersofmedia.hum.uva.nl/2007/10/07/repurposing-the-wikiscanner-comparing-dutch-universities-edits-on-wikipedia/ (accessed 12 January 2012).

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 the symposium *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, Oxford Internet Institute, pp. 1–17.

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.

DAGDELEN, D., FEUZ, M., ROOZE, M., POELL, T. & WELTEVREDE, E. (2010), 'Historical controversies now' [Online]. Available at: https://files.digitalmethods.net/var/historicalcontroversies/ (accessed 12 January 2012).

DANOWSKI, J. A. (2009). 'Network analysis of message content', *The Content Analysis Reader*, eds K. Krippendorff & M. Bock, Sage Publications, Thousand Oaks, pp. 421–430.

DIDIER, E. (2010) 'A theory of social consistency', paper presented at the workshop *After Markets: Researching Hybrid Arrangements*, Said Business School, University of Oxford, Oxford.

DUHEM, P. M. M. (1954) 'Physical theory and experiment', *The Aim and Structure of Physical Theory*, Princeton University Press, Princeton, pp. 180–218.

GROS, A. (2011). The economy of social data: exploring research ethics as device, Measure and Value, eds L. Adkins & C. Lury, Sociological Review Monograph Series 59, pp. 113–129.

HARAWAY, D. (1997) *Modest_Witness@Second_Millennium.FemaleMan©_Meets_Oncomouse™*, Routledge, New York and London.

HELMOND, A. (2010) 'On the evolution of methods: banditry and the volatility of methods', *Annehelmond.nl*, 17 May [Online]. Available at: http://www.annehelmond.nl/2010/05/17/on-the-evolution-of-methods-banditry-and-the-volatility-of-methods/ (accessed 12 January 2012).

KNORR-CETINA, K. (1981) *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*, Pergamon Press, Oxford.

LATOUR, B. (1987) *Science in Action: How to Follow Scientists and Engineers through Society*, Harvard University Press, Cambridge, MA.

LATOUR, B. (2005) *Reassembling the Social: An Introduction to Actor-Network-Theory*, Clarendon Lectures in Management Studies, Oxford University Press, Oxford.

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.

LEGGETTER, P. (2011) 'Real-time web or right-time web?', *Programmable Web*, 17 March [Online]. Available at: http://blog.programmableweb.com/2011/03/17/real-time-web-or-right-time-web/ (accessed 12 January 2012).

LEZAUN, J. (2007) 'A market of opinions: the political epistemology of focus groups', *The Sociological Review*, vol. 55, pp. 130–151.

LURY, C. (2012) 'Going live: towards an amphibious sociology', *Live Methods*, eds L. Back & N. Puwar, Sociological Review Monographs, Wiley-Blackwell, Hoboken, pp. 184–197.

MARKOFF, J. (2011) 'Government aims to build a ''data eye in the sky''', *New York Times*, 10 October, [Online]. Available at: https://www.nytimes.com/2011/10/11/science/11predict.html (accessed 11 January 2012).

MARRES, N. (2012) 'The redistribution of methods: on intervention in digital social research, broadly conceived', *Live Methods*, eds L. Back & N. Puwar, Sociological Review Monographs, Wiley-Blackwell, Hoboken, pp. 139–166.

MARRES, N. & ROGERS, R. (2005) 'Recipe for tracing the fate of issues and their publics on the web', *Making Things Public: Atmospheres of Democracy*, eds B. Latour & P. Weibel, MIT Press Cambridge, MA, pp. 922–935.

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

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*, Stanford University, Stanford.

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

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 [Online]. Available at: http://www.guardian.co.uk/media/interactive/2011/jul/13/news-of-the-world-phone-hacking-twitter (accessed 12 January 2012).

RIEDER, B. & ROHLE, T. (2012) 'Digital methods: five challenges', *Understanding Digital Humanities*, ed. D. Berry, Palgrave, Basingstoke, pp. 67–84.

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.

ROGERS, R. (2009) *The End of the Virtual: Digital Methods*, Amsterdam University Press, Amsterdam.

ROGERS, R. (forthcoming) *Digital Methods*, MIT Press, Cambridge, MA.

SARAWAGI, S. (2007) 'Information extraction', *Foundations and Trends in Databases*, vol. 1, no. 3, pp. 261–377.

SAVAGE, M., LAW, J. & RUPPERT, E. (2010) 'Digital devices: nine theses', *CRESC Working Paper Series. No. 86*, Manchester: University of Manchester.

SAVAGE, M. & BURROWS, R. (2007) 'The coming crisis of empirical sociology', *Sociology*, vol. 41, no. 5, pp. 885–899.

SHAH, C. & FILE, C. (2011) 'InfoExtractor – a tool for social media data mining', paper presented at the Third Annual Thematic, *Journal of Information Technology & Politics*. Conference, May 16 & 17, University of Washington, Seattle.

SLEE, T. (2011) 'Internet-Centrism 3 (of 3): tweeting the revolution (and conflict of interest)', 22 September [Online]. Available at: 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).

UPRICHARD, E. (2012) 'Being stuck in (live) time: the sticky sociological imagination', *Live Methods*, eds L. Back & N. Puwar, Sociological Review Monographs, Wiley-Blackwell, Hoboken, pp. 124–138.

UPRICHARD, R., BURROWS, R. & BYRNE, D. (2008) 'SPSS as an ''inscription device'': from causality to description?', *The Sociological Review*, vol. 56, no. 4, pp. 606–622.

WATTERS, A. (2011) 'Scraping, cleaning, and selling big data: infochimps execs discuss the challenges of data scraping,' 11 May, O'Reilly Radar [Online]. Available at: http://radar.oreilly.com/2011/05/data-scraping-infochimps.html (accessed 12 January 2012).

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.

WELTEVREDE, E. (forthcoming) 'Studying society, not Google? Repurposing Google for national Web research, in *Device Studies: Repurposing Web devices for social and cultural research*, PHD dissertation, University of Amsterdam.

WIKIPEDIA. (2012) 'Web scraping', last updated 10 January [Online]. Available at: http://en.wikipedia.org/wiki/Web_scraping (accessed 11 January 2012).

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

Noortje Marres Goldsmiths is Senior Lecturer in Sociology and Director of the Centre for the Study of Invention and Social Process (CSISP) at Goldsmiths, University of London. She studied sociology and philosophy of science and technology at the University of Amsterdam and was part of the team that developed the online research tool Issuecrawler. At Goldsmiths, Noortje convenes the MA/MSc Digital Sociology, leads an ESRC-funded project on Issue Mapping Online, and she has just published a monograph entitled *Material Participation: Technology, the Environment and Everyday Publics* (Palgrave, 2012). *Address*: Centre for the Study of Invention and Social Process, Goldsmiths College, University of London, London SE14 6NW, UK.

Esther Weltevrede is PhD candidate and lecturer at the New Media and Digital Culture program, University of Amsterdam. Esther is coordinating the Digital Methods Initiative and is also a member of Govcom.org. Her research interests include digital methods, device and software studies, national web studies, controversy mapping and the dynamics of online data. *Address*: Media Studies University of Amsterdam, Turfdraagsterpad 9, 1012 XT Amsterdam.