A reality check(list) for digital methods

Tommaso Venturini, Liliana Bounegru, Jonathan Gray, Richard Rogers

This is a preprint of: Venturini, T., Bounegru, L., Gray, J., & Rogers, R. (2018). A reality check(list) for digital methods. *New Media & Society*. http://doi.org/10.1177/1461444818769236

A Reality Check(-list) for Digital Methods

Tommaso Venturini
University of Lyon, Lyon (France)

Liliana Bounegru
Ghent University (Belgium); University of Groningen (Netherlands)

Jonathan Gray
King's College London (UK)

Richard Rogers
University of Amsterdam (Netherlands)

Big data, big infrastructures

Great expectations and great concerns have been raised about 'big data' (Boyd and Crawford, 2012), 'data science' (O'Neil and Schutt, 2013) and the 'computational social sciences' (Lazer et al., 2009). A rumbling storm of digital traces is said to loom over the humanities and the social sciences, bringing great power and also responsibilities. This prophecy of the 'data deluge' has some truth to it. In a handful of years, digital traceability has indeed bestowed our disciplines with larger and more diverse data sets than we have ever dreamt of. Yet, the deluge metaphor is also misleading as it mistakenly implies that the advent of social traceability is (1) unprecedented or (2) unproblematic.

First, while it is not wrong to emphasise the transformations brought to scientific knowledge by the growing availability of structured information, it is important to not forget that datafication was not born with digital technologies. In their book Big Data, Mayer-Schönberger and Cukier (2013) present dozens of examples of large and systematic campaigns of data collection ranging from census in ancient Egypt and China, to Renaissance bookkeeping, to 19th-century navigation.

Far from representing a break with the past, the traceability of digital media constitutes the latest development of the older phenomenon of 'media traceability'. Media are socio-technical systems that produce and enable inscriptions of individual and collective actions: 'media are our infrastructures of being, the habitats and materials through which we act and are' (Peters, 2015: 15). The specific way in which they enable (but also constrain) our actions is by translating them onto physical materials (stone, paper, copper, silicon, etc.). As old as cave painting, this process of inscription has drastically accelerated because of digital technologies:

> Once you can get information as bores, bytes, modem, sockets, cables and so on, you have actually a more material way of looking at what happens in Society. Virtual Society thus, is not a thing of the future, it's the materialisation, the traceability of Society. It renders visible because of the obsessive necessity of materialising information into cables, into data (Latour, 1998).

Not surprisingly, this growing traceability of collective actions has affected social sciences. Researchers have always relied on media inscriptions to investigate collective phenomena, and the advent of digital media has increased the quantity and variety of the traces at their disposal (Venturini et al., 2017). Hence the understandable excitement of social scientists finally having access to data sets that are as large and rich as those of their colleagues in the natural sciences (Venturini et al., 2015).

Second and as a constraint to this excitement, the increase in the information available on social phenomena does not come for free. Digital traceability may provide the social sciences with *quantities* of information that are comparable to those collected in natural science laboratories, but the *quality* of such traces is radically different. Unlike the traces produced by a telescope or a microscope, media inscriptions are (in general) not created *by* or *for* the academic community (Burrows and Savage, 2014; Marres, 2017; Savage and Burrows, 2007).

Although both the Internet and the Web were initially incubated in universities and research institutions, those times have passed (with the partial exception of the Internet Archive and Wikipedia). Nowadays, digital media belong to

the companies and institutions that have paid the huge costs necessary to set up their infrastructures (Frischmann, 2001). This is true at every level of digital networks: from the submarine cables bringing Internet to every corner of the planet (Boullier, 2013), to the social platforms allowing anyone to push and pull information from the Web. Digital media have not developed by themselves. They *have been developed* by the investments of a (limited) number of public and private organisations which are now the gatekeepers of their traceability. The metaphor of 'media ecologies' (Fuller, 2005), often used to describe the interactions within and among media, is in this sense improper. Media are not natural ecosystems evolving spontaneously out of 'chance and necessity' (Monod, 1970). Instead, they may be understood as artificial 'landscapes' (Rogers, 1999), carefully cultivated by the actors that participate in their construction (Rein and Venturini, 2018).

However, the fact that digital inscriptions are created outside academia does not make them unfit to be used for research purposes. On the contrary, the growing dominance of these platforms makes the academic exploration of their traces even more important (Burrows and Savage, 2014; Savage and Burrows, 2007). And, while researchers might not harvest all media inscriptions without the collaboration of the infrastructures that created them, they can still obtain partial access to the traces. It is a feature of the political economy of contemporary media that information is not only accumulated but also partly redistributed. Media companies collect information from us and redistribute it in various configurations and products as part of their business strategy (see, for example, Bodle, 2011). Google, Facebook, Twitter and the likes may strive to collect and monetise our messages, clicks and hyperlinks, but in doing so, they also provide us with insights from the data that they collect. This is not simply a compensatory move: it is part of the strategies for platform to present themselves as providers of valuable analytics and partners to established and emerging data industries.

While digital platforms still rely on classic business models based on the exchange of information against money (through subscriptions) or against attention (through advertising), they are also increasingly trading information against other information. Every time you use its search engine you provide Google information about your interests, but in exchange you are given access to the largest database ever created, through the most advanced information engine ever conceived. And, although (according to the *terms of service*) you are not supposed 'access [the information] using a method other than the interface and the instructions that we [Google] provide', nothing prevents you from being methodologically inventive and re-using this information for purposes other than those which were originally intended (Lury and Wakeford, 2012). The same applies for most Web and mobile services, whose business model implies some redistribution of data.

The 'digital methods' approach

Creatively repurposing the traces and the methods inscribed in digital objects is the aim of an emerging approach that has come to be known as *digital methods* (Rogers, 2013). This approach aims at exploring 'how may one learn from how online devices (e.g., engines and recommendation systems) make use of the objects, and how may such uses be repurposed for social and cultural research?' (Rogers, 2009: 1).

Precisely because they propose to do 'research *with* the Internet' – that is to say by exploiting the information made available by Internet platforms – digital methods are not suitable for all research scenarios. For this approach to make sense, the investigated phenomenon must be to some extent performed or, at least, reflected in such platforms, and this is clearly not the case for all collective events. As far as the digital tendrils may have extended, there still are plenty of crucial social dynamics that play out prevalently in face-to-face interactions (or through non-digital media). These include, of course, interpersonal exchanges taking place in homes, classrooms and offices, and also (and in a way that is often forgotten) the very situations in which digital messages are received. The non-digital situations in which digital media are consumed (at work, on the couch, in an Internet café, on the subway through a smartphone, etc.) and the ways in which their contents are processed through direct conversations may influence crucially their reception (as repeatedly shown in the case of traditional broadcast – cf. Gans, 1993; Katz, 2001; Staiger, 2005). Yet, these influences remain outside the grasp of digital methods and should be assessed through other techniques.

In addition, while digital methods may, in theory, apply to the inscriptions produced by any digital infrastructure, most studies following this approach focus on the largest online platforms. Engaging vast and non-specialised

populations, services such as Google, Facebook, Twitter and Wikipedia have understandably captured the interest of computational social scientists. As a consequence, while many tools exist to investigate the giants of the web, smaller and specialised platforms remain relatively unexplored (Venturini et al., 2014).

Besides these general limitations and even when scholars investigate phenomena clearly visible in mainstream online platforms, caution is required. Repurposing the interfaces, databases and methods of digital media may be extremely useful, but it demands some vigilance. To produce useful and interesting findings, digital methods require the extra care needed for the secondary analysis of inscriptions that have not been created by or for the social sciences and thus bear the *imprint* of the particular purposes (whether political, commercial or otherwise) and technical infrastructures through which they were created. Using digital methods, we are always at risk of mistaking the characteristics of medium for the *signature* of the phenomena we wish to observe.

To a great extent, this cannot be helped. Since the work by McLuhan (1964; McLuhan and Fiore, 1967), we know that 'the medium is the message' and that electronic infrastructures do not simply transport social phenomena but also participate in their production. This basic constructivist acknowledgement, however, should not be taken as an excuse for careless work. It is precisely because there is a priori no clear separation between noise and information, that efforts should be invested to distinguish them a posteriori (Marres and Moats, 2015). This operation of accounting for the entanglement of the digital devices and the actions which they constrain and enable (contrasting the specific signal of the observed phenomenon to the general background of the repurposed medium) is vital, but often overlooked.

In this article, we provide a basic list of precautions which may be taken when using digital methods. Others have already discussed various 'perils' (Bollier and Firestone, 2010), 'provocations' (Boyd and Crawford, 2011), 'challenges' (Rieder and Röhle, 2012), 'problems' (Marcus and Davis, 2014), 'traps' (Lazer et al., 2014) and 'misunderstandings' (Venturini et al., 2014) of digital research. While these and other inquiries offer interesting reflections about the way in which digital technologies affect the epistemology of social sciences, this article has a more *practical* objective: it aims at collating a series of caveats that scholars may bear in mind while designing their digital research.

Since we are particularly interested in the interferences between the production of inscriptions by digital media and their repurposing by scholars, we will focus on the first phases of digital research, where the entanglements between technological infrastructures and social phenomenon are stronger. With reference to the four 'analytical moments' identified by Mikkel Flyverbom and Anders Koed Madsen (2015) – production, structuring, distribution and visualisation – we concentrate on the first two stages of research (but some discussion of the latter can found in Venturini et al., 2015 and in Gray et al., 2016).We will illustrate our precautions with a series of studies developed at the Digital Methods Initiative (DMI) at the University of Amsterdam (digitalmethods.net). Although equally interesting projects have been carried out in many other research institutions, we decided to focus on the DMI to facilitate the comparison. In addition, the research carried out at DMI has the advantage of being accompanied by web pages presenting the research methods, protocols and data sets (cf. wiki.digitalmethods.net/Dmi/DmiSummerSchool and wiki.digitalmethods.net/Dmi/WinterSchool, but specific URLs will also be provided).

A few preliminary definitions

For the sake of clarity, we will formulate our precautions as four sets of questions (from the most theoretical to the most practical). This arrangement is to a large extent artificial: in the practice of digital research, these precautions overlap and problems have the unruly tendency to arise in knots rather than in orderly sequences.

Before moving to the checklist, we will introduce some of the key notions used in this article. The descriptions we provide below should not be understood as strict definitions but as a working heuristic for digital researchers:

> A *medium* is any technical infrastructure that allows the organisation and extension of collective actions in space or time. The printing press, television, telephone and Web are media in which they allow social actors to interact without being in the physical presence of each other. One should be not fooled by this simple definition into

taking media infrastructures for granted. Classic works in media studies emphasise how media are not neutral agents and instead play an active role in the articulation of meaning and communications. For example, Howard Innis' pioneering studies (1986, 2008) highlighted what he called the particular characteristics or 'biases' of media and how they enabled different social institutions of law, religion, culture and commerce. Drawing on Innis' work, Marshall McLuhan contributed to the recognition of media systems as objects and sites of study. James Carey recognises Innis and McLuhan's role in establishing the study of media as 'not merely [...] appurtenances to society but as crucial determinants to the social fabric' (1967: 270–271). He underlines the limits of models focussing on 'transportation and transmission' and instead proposed to consider media as processes through which 'reality is produced, maintained, repaired, and transformed' (2009: 19).

A *platform* is a specific way of organising a media infrastructure, not only constraining the way in which the medium can be employed but also facilitating its exploitation. Recent research has focused both on the rhetorical aspects of platforms (Gillespie, 2010) and on their 'material-technical' characteristics (Helmond, 2015). Facebook is an interesting example of how a limited repertoire of 'sentiments' – including 'likes' and a number of 'reaction' emoticons – are facilitated through a *centralised* social media company and then extended in relation to almost any digital media.

A *scientific inscription* is any piece of information that is materialised through the use of a technical device for the purposes of research. Inscriptions are the foundation of any scientific enterprise for they allow to imprint knowledge on materials which can be stored, transformed and transmitted (Latour, 1985; Latour and Woolgar, 1979).

A *digital trace* is any inscription produced by a digital medium in its mediation of collective actions – for instance, a post published on a blog and a hyperlink connecting two websites or the log of an e-commerce transaction. We call this particular type of inscriptions 'traces' as a reminder that they are (most often) generated by purposes other than academic research. Some of these inscriptions are 'native' to digital media, while others are originally analogue and digitised at a later stage.

A *corpus* is an ensemble of inscriptions or traces that have undergone the process of selection, cleaning and refining necessary to prepare them for scientific analysis. For instance, hyperlinks are a classic example of digital traces, but they only become a research corpus when they are translated into constricted lists or into arcs connecting a network of websites.

The notion of *digital methods* was introduced in 2007 as a counterpoint to virtual methods, which sought to introduce the social scientific instrumentarium to digital research (Rogers, 2009). Virtual methods, it was claimed, consisted in the digitisation of such traditional research methods (e.g. in online surveys or online ethnography). Rooted in media studies and the so-called computational turn in the humanities and social sciences, digital methods sought instead to learn from the methods of the medium and repurpose them for social and cultural research. Reflecting on 'natively digital' methods sensitised the researcher to the specificities of the then 'new media', to their effects, platform vernaculars and user cultures. 'Following the medium' also would offer the researcher a strategy to cope with the ephemerality and instability of the Web, where a new feature, a changed setting or the shutting down of an Application Programming Interface (API) could stymie longitudinal studies. While remaining critical of the implications of such changes, digital methods would ask which kind of research the platform affords. Digital methods, thus, may be defined as techniques for the ongoing research on the affordances of online media.

1. Digital media and objects of study

The first checkpoint encountered in all digital methods projects concerns the *adequacy* of the source exploited in relation to the object of the study. The adequacy can be defined as the *extent* to which the observed phenomenon takes place within the medium that is repurposed to examine it. If one is interested in the public of a particular issue, one might ask whether and how this public is online, whether its members use a given platform, space or device from which digital data are collected (Twitter, Facebook, a website or a mobile app), and what kinds of technical skills, capacities and networks they have available to them (cf. Hargittai and Hsieh, 2013).

Say, for example, that you are investigating data collected through the Steam gaming platform (steampowered.com). Different cautions will be needed depending on the ambition of your research: Do you plan to describe the gaming habits of Steam's users? Or are you interested in online gaming trends? Or do you want to inspect the cognitive effects of computer games? Or question the social role of game playing in general? If you are studying the practices of specific platform (e.g. the habits of Steam gamers), then the inscriptions produced by that platform constitute the *primary* traces of the phenomenon you are after. But if you use those particular activities as an example of a more general phenomenon (e.g. collective game playing), then your traces will only offer you a partial observation.

As should be clear from the example, the distinction is neither binary nor written in stone. It depends on how you define the scope your investigation and can change as your research evolves. It should also be noted that working with partial traces is not necessarily an insurmountable problem. If it is true that the larger is the coverage of your study object, the easier it will be to generalise your findings, it is also true that the more phenomena and media coincide, the harder it is to separate them analytically. In the paragraphs above, we have used the expression 'to take place in'. While this expression conveniently describes the way in which actions happen within or beyond a specific medium, it has the disadvantage of artificially separating collective actions from the medium that supports them. Media are not only 'places', 'spaces' or 'contexts' but actors themselves whose actions interfere and transform the behaviour of their users (Castells, 2009). These 'media effects' should be taken into consideration to understand that the phenomena we observe are not just *hosted* and *traced* by the media in which they occur, but are also deeply shaped by them.

1.1 How much of your study object occurs in the medium you are studying?

In its simplest definition, a collective phenomenon can be defined as a network of interfering actions (Latour, 2005). These actions can be of very different kinds, varying from an occasional intervention of an individual actor (e.g. when a customer makes a bid in an online auction) to a long-standing configuration of socio-technical forces (e.g. the legal constraints implemented in the mechanism of the bidding interface). What counts here is the extent to which the actions that comprise the phenomenon you wish to observe are mediated by – and, therefore, leave traces in – the medium that you are repurposing.

If you are studying learning practices in Massive Open Online Courses (MOOCs), you may, for example, safely assume that most interactions that you investigate may take place through the MOOC platform and therefore be recorded by it. But if you are studying the life of a university through the records of its administrative systems, you should be aware that most of the informal face-to-face exchanges that constitute a crucial part of college experience will not show up in your data set.

A good example of a close alignment between the research object and the medium is a 2016 study of conflict resolution practices in Wikipedia (Weltevrede and Borra, 2016). This study draws on a project which was initially meant to use Wikipedia's traces to identify emerging societal controversies (contropedia.net) (Figure 1). Soon, however, it became clear that while tensions coming from external debate are often mirrored in the online encyclopaedia, such conflicts are hard to distinguish from the internal quarrels around the platform's architecture, policies and guidelines. Acknowledging this difficulty, the study shifted the focus of its inquiry to examine practices of coordination specific to the platform and the distinctive ways in which they facilitate collaboration and defuse conflict.

This is a preprint of: Venturini, T., Bounegru, L., Gray, J., & Rogers, R. (2018). A reality check(list) for digital methods. *New Media & Society*. http://doi.org/10.1177/1461444818769236

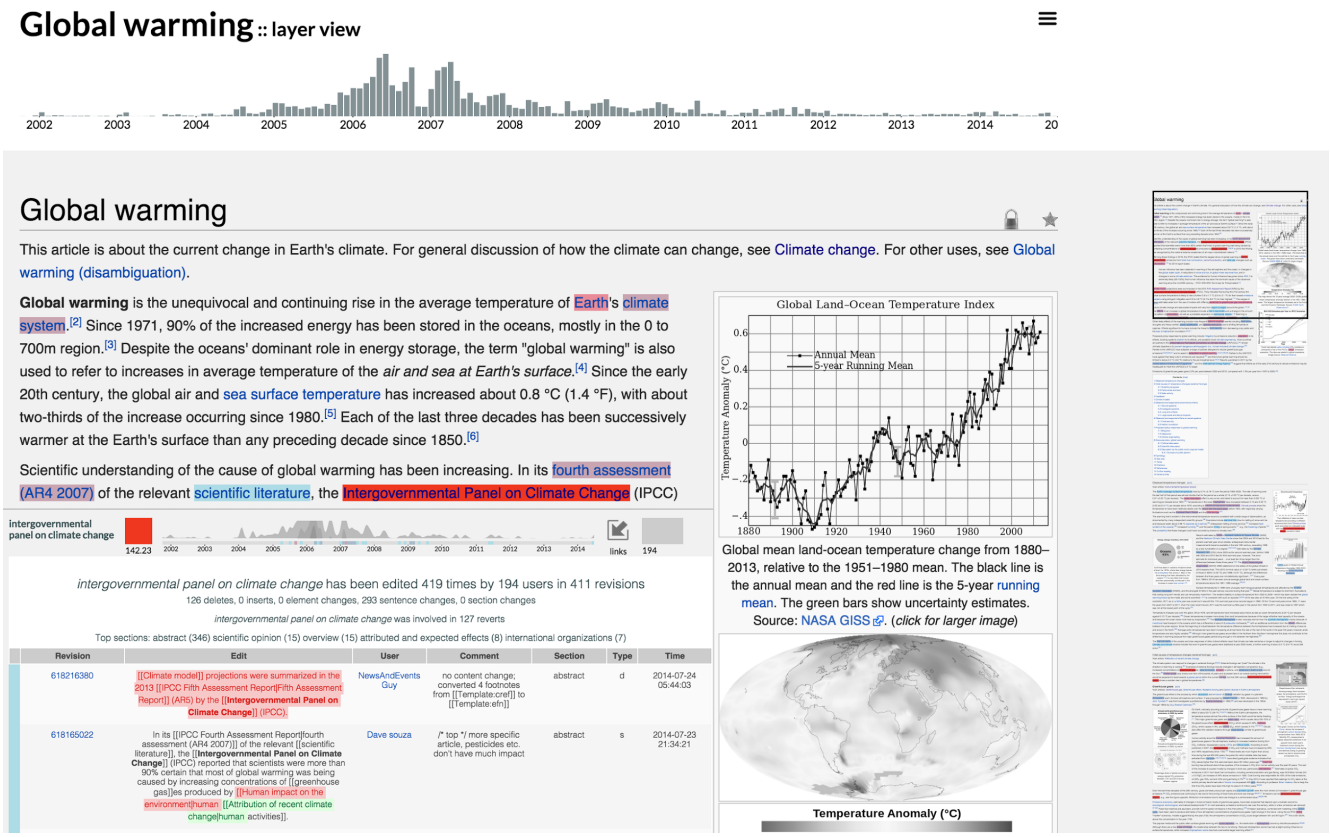

Figure 1. Two screenshots of the Contropedia.net interface. Controversial wiki-links are highlighted in red on the original page and the full evolution of the discussion surrounding them is displayed below (original figure in Weltevrede & Borra, 2016).

Other times, the partiality of medium coverage with respect to the phenomenon may be used strategically. Drawing on James Gibson's (1986) theory of visual perception, Anders Koed Madsen (2012, 2015) introduced the term 'web-vision analysis' precisely to point at the way in which researchers can use different media and filtering parameters to compare different angles on the same phenomenon:

> Web-visions are cases that result from deliberate combinations of devices and tools, and the mode of seeing that results from these combinations is the basis of their potential relevance ... the researcher is left with an arsenal of variables that can be used to manipulate the construction of the web-visions in a quasi-experimental fashion. The mode of seeing can, for example, be tweaked by altering the logic of filtering in the delineation device, the country of origin of the device, the language used to query the device or the settings of the web-crawler used to construct the visualization. (Madsen, 2012: 62)

Partiality, in other words, is not always a liability. Purposefully moving away from the main site where the phenomenon occurs and where it is typically studied may offer fresh angles and perspectives.

1.2. Are you studying media traces for themselves or as proxies?

A subtler dimension of the question of coverage has to do with the *nature* of the actions traced in the medium that you are investigating. Do they constitute the very phenomenon that you are examining or are they the occasion to study *other* actions not directly traced in the data at your disposal?

Social media, for example, have become a mine of information for the study of social movements as civic organisations increasingly rely on them to coordinate the actions of their members both online and offline (Gerbaudo, 2012). Yet, it is one thing to use traces from social platforms to investigate online mobilisation and another to use them to study street protests (Rogers & Marres, 2002). In the former case, the messages exchanged online constitute the very object of the study, in the second they are the proxies of other actions (walking, standing,

shouting…) taking place outside the medium. Indeed, digital methods takes the explicit stance of using digital traces to study not only online phenomena but culture and society in general (Rogers, 2013, 2017). Repurposing the media means using digital traces as proxies for phenomena that extend beyond them.

In an exploratory project, for example, a group of researchers compared the Google Web Search results for the query "rights" in a number of languages, to highlight the specific ways in which cultures conceive the question of human rights (Bekema et al., 2009; https://www.digitalmethods.net/Dmi/NationalityofIssues; Figure 2).

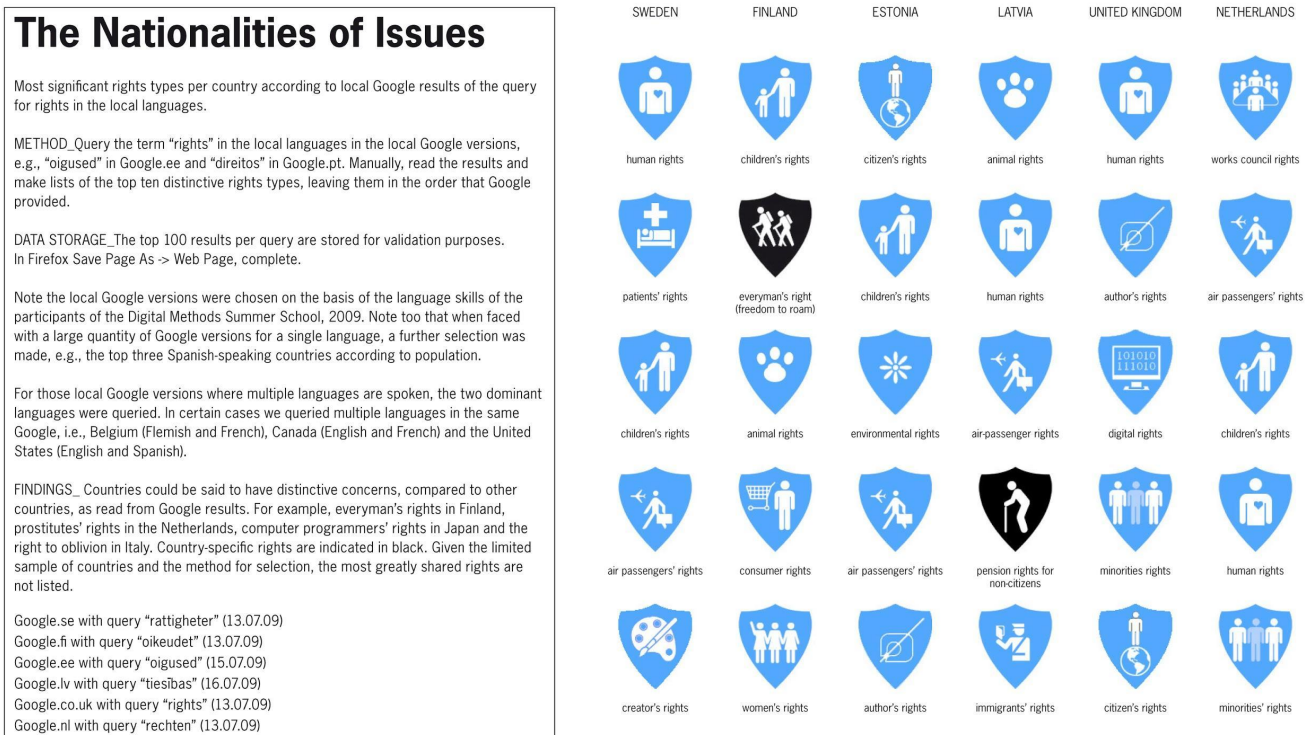

Figure 2. A visual representation of the different human rights as appearing in the results of Google Search for different countries and languages (original figure in Bekema et al., 2009).

2 Definition of the object of study

As we have just seen, the key to securing the adequacy between observed phenomenon and repurposed medium is to handle with care the relation between the scope of your research questions and the traces that you will use to investigate them. In the previous point, we considered such questions 'passively' as if the only thing researchers could do is to choose a source that fit their ambitions. Yet, researchers can also (and in fact *should* also) actively and creatively operate to align the two. This process is called 'operationalisation' and it refers to the way in which the entities that you wish to observe are defined through the traces at your disposal (see, for example, Moretti, 2013). In digital methods research, this takes the shape of 'an on-going process of assembling, re-configuring, and aligning research questions with digital media and device cultures' (Weltevrede, 2016).

Suppose you want to observe the connections between private companies, public institutions and civic groups through the way in which they refer to each other in their online discourse. There are a number of different ways in which you can operationalise this research question. You can not only look at the hyperlink network among the websites of your actors, but you can also consider the overlap of their Facebook friends. You can follow the retweeting of their representatives, or explore the connections among their pages on Wikipedia. All these operational definitions are legitimate, but each of them will give you a different view on your object of study with different possible biases that should be carefully considered.

Even within a single platform, different operational definitions of the same research object are often possible. Take the case of the investigation of controversies in Wikipedia. Because of the way in which MediaWiki (the software that supports the famous collaborative encyclopaedia) stores information, 'controversiality' can be operationalised not only at the article level (to highlight which topics are disputed) but also at the level of smaller elements such as the links within the articles (to reveal, for instance, which references are most contested). In addition to this, multiple measures of controversiality may be defined, from the volume of edit histories, to the depth of discussions in associated talk pages (Borra et al., 2014; Weltevrede and Borra, 2016). Each of these operationalisations leads to a different appraisal of what constitutes a matter of concern or an expression of disagreement.

2.1. Is your operationalization attuned to the medium formats?

To validate your operationalisation, start by considering its agreement with the source in which it will be deployed. Working with secondary data, you do not have the leisure to define your objects of study as you wish, but you are obliged to consider (at least in part) the way in which they are formatted by the technical and organisational standards of the medium.

For example, when investigating public debate through Twitter, one cannot avoid acknowledging that topics in this platform tend to be organised through a very specific technical object: the hashtag. This object has distinctive features. It is always preceded by the '#' symbol, it acts as a topical marker, it assembles publics around a shared matter of concern and it can be used as a keyword for monitoring or searching content. These features influence the way in which actors discuss and also the manner in which research can investigate such discussions (Marres and Gerlitz, 2016).

Let's say you want to use Twitter to explore the groups engaged with the issue of public finances. Following a traditional sociological approach, you may profile publics according to geography, demographic features or societal sectors (see, for example, McCormick et al., 2015; Sloan et al., 2015). But if you want to ensure that your line of inquiry is attuned to Twitter's practices, these might not be the best starting points. Twitter follows a different approach to knowledge production than classical 'research devices'– through follower–followee relations, liking, linking, tagging and curated 'moments' (see Ruppert et al., 2013). In a project on the dynamics of European public finances, we, thus, focused on the specific forms of engagement facilitated by Twitter. For example, we collected data about a series of hashtags (e.g. EUBudget, 'EU budget' or #ESIF) and explored the associated actors and issues (e.g. #refugeecrisis, #youth, #OurFundsOurRights, #Regeneration and #Brexit). Through such analysis, we observed the formation of new publics, as well as dynamics of 'hashtag hijacking' – the convergence of different social worlds through the accidental or purposive use of similar key words.

2.2. Is your operationalization attuned to the medium practices?

The technical implementation of actions and actors, however, is only one of the ways in which digital mediation structures your research object. Another one has to do with the practices employed by the users of the medium. While the technical infrastructures clearly influence the interactions that take place through them, they are also open to interpretation (see Gillespie et al., 2014, Paßmann and Gerlitz, 2014). Uses and technical formats are not independent – actors both 'do with' media affordances and influence the way in which such affordances evolve (Bucher and Helmond, 2017). For example, while Twitter offers an official way to signal association between different accounts, through the 'follow' function, such associations have been proved to be weaker than the action of 'mentioning' or 'retweeting', both of which have been initiated by users and only later officially adopted by the platform (Kooti et al., 2012).

Your operationalisation, therefore, should be adjusted not only to the technical infrastructure of your medium but also to the practices of its users. In anthropology, this question is addressed through the distinction between 'etic categories' (the concepts employed by the researchers and their peers) and 'emic categories' (the notions employed by the community under research) (see Munk, 2013, for a discussion of how some classic anthropological notions

can be applied to digital phenomena). This distinction reminds us that the intellectual tools that we use to describe a collective phenomenon should be respectful of the way in which the actors conceive their own social existence.

This care is particularly crucial as 'query design' is concerned (Rogers, 2017). The ways in which actors label the phenomena in which they are engaged can be subtle and complicated. For example, one may note that climate 'scepticism' is the self-description preferred by those who doubt the human causes of climate change, while climate change 'denial' is the notion used by their opponents (Niederer, 2013). Understanding the nuances of emic language can help you capture different sides of your object and the competitive ways in which different groups frame the same phenomena (see the concept of 'equivalence framing' in Cacciatore et al., 2016). It also allows you to generate better and more precise queries. A recent study of 'mental illness' on Tumblr, for example, started from the generic query #mentalillness (Sanchez-Querubin et al., 2016). Soon, however, the co-hashtag network around this query (the hashtags most often used alongside #mentalillness) revealed that the concept of #recovery characterises the most significant practices associated with mental illness on Tumblr and thus became the focus of the study.

3. From single-platform to cross-platform analysis

So far, we have discussed the cautions necessary to handle digital traces, under the assumption that those traces derived from one medium. Of course, this is not always the case. Most collective phenomena tend to extend beyond the frontiers of any single medium and often force researchers to follow them across different media. Cross-platform projects tend to be richer than single-platform ones, as they allow to compare the findings observed in one medium with those obtained in others. Depending on your research question, this 'triangulation' can help to separate the characteristics of collective phenomena from the features of the media.

For example, when observing the rapidity with which issues rise and fall on Facebook, it is difficult to decide whether such rhythm is an indication of a superficial debate or an effect of the platform which encourages shorter attention spans. Most probably, both are true. However, by comparing how the same topics evolve on Facebook, in the blogosphere, in newspapers or in the scientific literature, one can distinguish the underlying features of a collective phenomenon from the specific way in which it is enacted in a particular medium.

3.1. Does your study object spill across several media?

Going 'cross-platform' is indispensable when the media themselves encourage the circulation of the same contents across their borders. This is often the case in Web-based media; for hypertext protocols facilitate the creation of connections. Despite all the discussion about the 'walled-gardens' of social media, every platform is connected to other platforms and sometimes to different media.

Twitter is particularly illustrative in this sense. Given the word limit of its messages, this platform has from its inception been used as a device to point to contents published elsewhere. Building on such characteristic, most other social platforms offer functionalities of 'automatic tweeting'. As a consequence, Twitter's dialogues are often influenced by the echoes of the discussion happening in other contexts (Gerlitz and Rieder, in press).

Studying the circulation of fake news, for example, we soon realised the impossibility to limit our study to a single medium (Bounegru et al., 2018 – fakenews.publicdatalab.org). The danger of successful fake news stories comes less from their falseness (which is in most cases easy to detect) than from the virality with which they bounce from a medium to the other and thereby steadily occupy the public agenda (see also Leskovec et al., 2009; Shifman, 2013) (Figure 3).

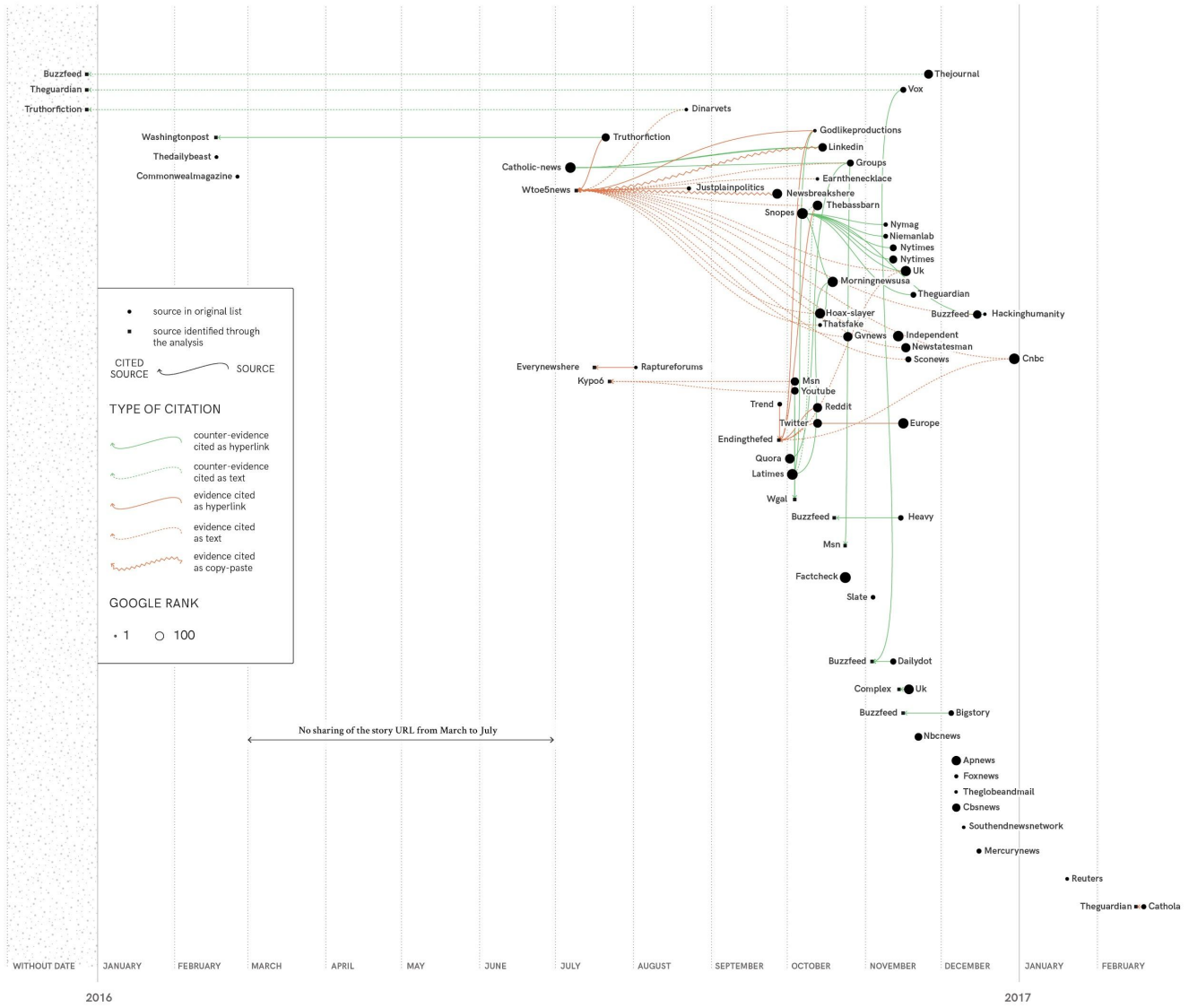

Figure 3. Spread and debunk of the fake story according to which the Pope would have endorsed Donald Trump. The nodes represent the web pages in which the story has circulated and the lines the different ways in which they mention each other (original figure in Bounegru et al., 2018).

3.2. Do you use different but comparable operationalisations for different media?

While 'cross-platform' research can be useful and sometimes indispensable, it also entails additional difficulties due to the necessity of developing multiple operational definitions of the entities under consideration. Each of these definitions should not only be attuned to the specific medium in which they are used but also be sufficiently consistent to allow comparisons.

Earlier, we considered the investigation of how actors associate online. The most straightforward choice would be to operationalise their connections as the hyperlinks connecting their different online *personae* (their website, their Facebook page, their Twitter account, etc.), as the notion of hyperlink is defined at a low layer of Web protocols and does not depend on the specific platform implementation. Such choice, however, would fail to recognise that different platforms and practices. Hyperlinks among websites may better be translated by 'friendships' or 'likes' in Facebook and by 'retweets' and 'mentions' in Twitter. In cross-platform approaches the trade-off between

attunement and comparability is always problematic and one should find specific solutions coherent with the aims and the constraints of the research.

In an ongoing study, we compared different media to reveal competing framing of open data politics (Gray et al., 2016). On Twitter, many actors seemed to cluster around topics related to business opportunities (such as #startup, #smartcity or #innovation) as well as transparency and open government (#ogd, #opengovernment, #transparencia). By contrast, by analysing the Wikipedia pages connected to the theme of open data, we observed topics such as 'open source software', 'free software movement', 'open access', 'free culture movement' and 'Creative Commons' – indicating how open data are articulated less as a policy or economic issue and more as part of the 'digital commons' movement. Finally, newspaper analysis suggests that open data are frequently discussed in relation to international development.

4. Corpus demarcation and data access

Once you have chosen your object of study, the media through which you will examine it and how to operationalise passage from the one to the others, you are still confronted with practical difficulties. We have gathered them in this final checkpoint because they concern the way in which media inscriptions are turned into a research corpus.

For the sake of simplicity, we have so far discussed the adequacy between digital traces and research ambitions considering vast social phenomena (e.g. the sharing economy of housing) and entire media platforms (e.g. Airbnb). Such a breadth of scope is, however, likely to be highly inadvisable in many cases as it may only yield superficial results. Instead, researchers should concentrate on specialised questions (e.g. whether peer-to-peer renting has professionalised in a given city and in a given and period of time) and on restricted subsets of the traces generated by the medium they investigate. The necessity of *selecting* a specific object of study from the fabric of collective life and a correspondingly delimited corpus out of the web of digital traces is a crucial operation and one that raises a few delicate questions.

4.1. What does your corpus represent?

In the question 'How much of your study object occurs in the medium you are studying?', we discussed how the traces offered by one medium are rarely co-extensive to the research ambitions. The problem of partial coverage is intensified by the fact that the data of any given research are always a subset of the traces offered by their source. Every time you use a query (or a set of queries) to extract information from a platform, you should infer which words are used by the actors you are interested in to define their matters of concern. Sometime this inference is straightforward – commercial brands, for example, invest huge efforts to standardise their name – but think of how many different expressions can be used to refer to 'environmental degradation'. How to be sure that the query you use results in sufficient coverage of your study object?

The problem here is not to be exhaustive. Exhaustiveness is a false ideal in digital research – not only because there are just too many digital traces out there for researchers to hope to seize them all but also – and more importantly – because extending one's coverage may produce more noise than signal. When we say that media traces are not co-extensive with social phenomena, we mean that the former are narrower *but also* broader than the latter. The blogosphere, for instance, does not contain the climate debate because such debate also occurs in many other media (scientific literature, news, social platforms, etc.), but the reverse is also true: climate change is only one of the innumerable topics discussed online.

Digital corpora can never be exhaustive. They can, however, be representative. They should not necessarily cover *each and every* thread that constitutes the fabric of social phenomena, but they should not tear such fabric or artificially extend it. The notion of representativeness may be inappropriate here, for it is associated with clear statistical definitions that are inapplicable to digital methods. For most digital research, there is no straightforward statistical test to assess the validity of a corpus. The best you can do is to describe explicitly the various operations of selection and transformation that connects the original traces to the final corpus and reflect on their analytical consequences.

4.2. Are you accounting for the ways in which data are 'given' by the media?

Much has been written about the unfortunate etymology of the word 'data', which conveys the false impression that information is *objectively given* and not constructed not only by the researchers but also by the technical infrastructures that have generated those data, their users and the companies and institutions that own those infrastructures (Bowker, 2013; Drucker, 2011). Acknowledging that we receive our information from someone else (data are 'given' at least in this sense) brings our attentions to the conditions of such *delivery*.

The sources from which we derive our inscriptions and the instruments through which we acquire them have consequences on the quality of our observations. When observed through the traces that it leaves on Twitter, public debate often appears as a chaotic flux of conversations ephemerally agglutinating around emerging ideas, while struggles between overarching world visions and systems of forces become almost invisible. As the saying goes in digital methods community, 'when all you have is a Twitter feed, everything looks like a hashtag'. Electronic media do not merely record the interactions that they mediate – not unlike social researchers, they also measure and analyse (Marres, 2012). They count them beside making them countable (Agre, 1994; Gerlitz and Rieder, in press).

Investigating climate debate on Twitter, Marres and Gerlitz (2016) noted that the platform relies on 'frequency of mentions' to identify and promote trending topics. Such focus encourages specific practices among the users (e.g. retweeting as way of having messages picked up by the system) and is transmitted to most Twitter analytic tools (Figure 4). This ends up privileging hashtags referring to events or campaigns (e.g. #cop16, #auspol and #savethearctic) that are subject to hype-like dynamics. In order to detect more substantial issues, the researchers then moved from frequency measures to 'associationist measures' (not how many times a hashtag is mentioned, but how many other hashtags co-occur with it), which allowed them to identify tags such as #economics, #flood, #co2, #health, #environment and #drought.

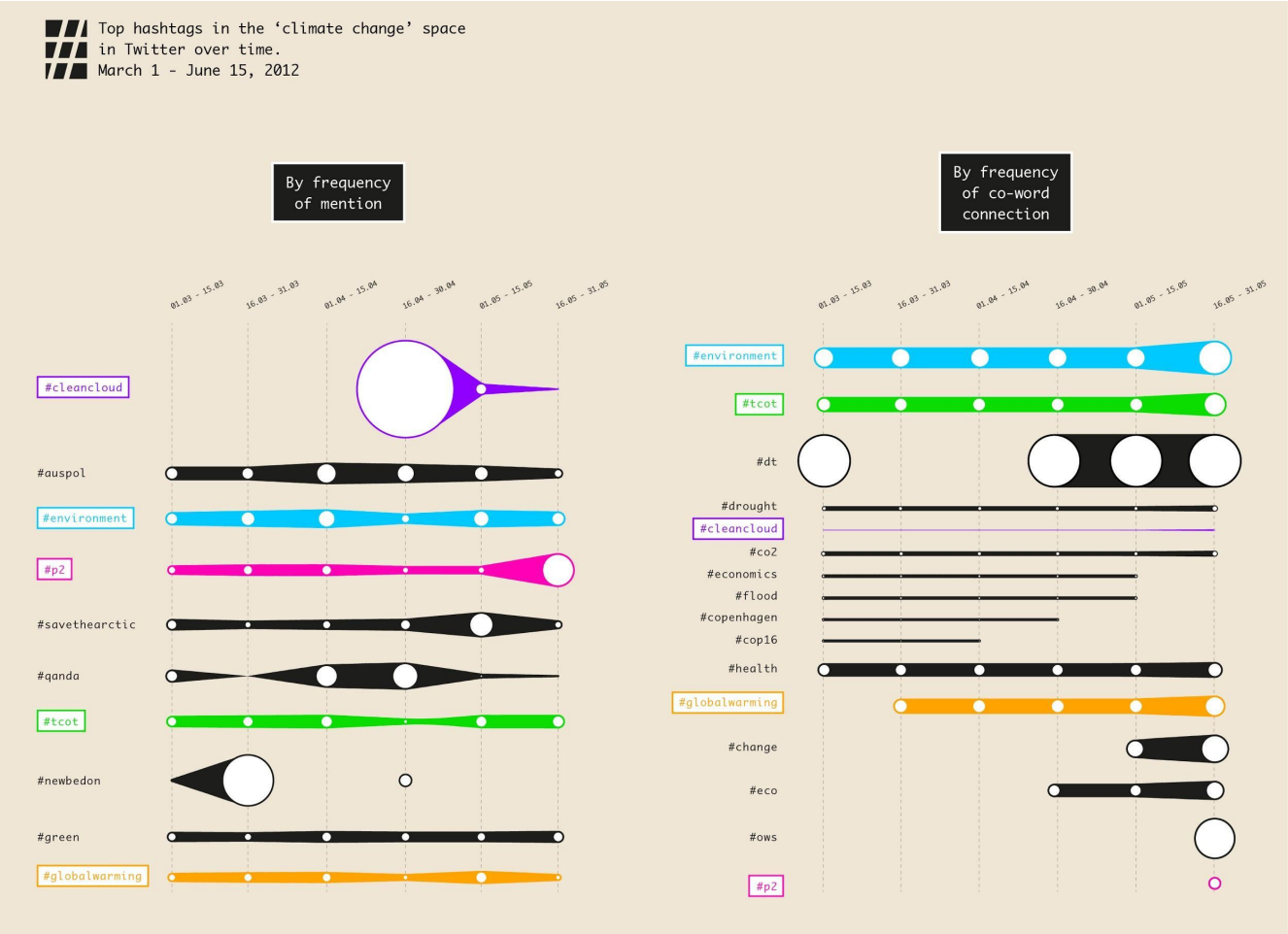

Figure 4. Comparison of the most mentioned and most connected hashtags connected to climate change debate in Twitter (original figure in Marres & Gerlitz, 2015).

Not only different digital traces are infused with the technical, commercial and ideological premises of the platforms that generate them (cf. Havens and Lotz, 2012; Mandiberg, 2012; Srnicek, 2017), but our data sets depends on our entry point to digital inscriptions. For example, most digital platforms provide an API that structures what and how much information may be accessed, as well as by whom and with which restrictions. The information accessed through such 'pipelines' is often significantly different in detail and completeness from that displayed on the interface of the same platforms (as a result of operations of aggregation, anonymisation or normalisation). Sometimes important portions of digital traces are excluded from APIs – the Facebook API, for example, recently withdrew all information on personal profiles due to privacy requirements, although such profiles constitute the bulk of Facebook's inscriptions (Rieder, 2013). The possibility remains, of course, to 'scrape' information directly from the publicly accessible interfaces, services and applications, but even in this way traces bring with them the mark of their origin (Marres and Weltevrede, 2013).

Conclusion

Instead of concluding with a theoretical discussion, we prefer to remain faithful to the practical approach of this article and provide a summary of the eight questions discussed above. This summary is offered in the form of an *aide-mémoire* that researchers embarking upon digital method projects can keep with them as a reality checklist of their findings and interpretations:

1. **Role of digital media in relation to object of study**
    - 1.1. How much of your object of study occurs in the medium that you are studying?
    - 1.2. Are you studying media traces for themselves or as proxies?
2. **Definition of the study object**
    - 2.1 Is your operationalisation attuned to the formats of the medium?
    - 2.2. Is your operationalisation attuned to the practices of the medium users?
3. **From single-platform to cross-platform analysis**
    - 3.1. Does the phenomenon that you are studying spill across several media?
    - 3.2. Have you different but comparable operationalizations, for the different media?
4. **Corpus demarcation and data access**
    - 4.1 What does your corpus represent?
    - 4.2 Are you accounting for the ways in which data are 'given' by the media?

References

Agre PE (1994) Surveillance and capture: Two models of privacy. *The Information Society*, 10(2), 101–127. Available at: http://doi.org/10.1080/01972243.1994.9960162

Bekema V, Bounegru L, Fiore A, et al. (2009) Nationality of issues: rights types. *Digital Methods Summer School.* Available at: https://www.digitalmethods.net/Dmi/NationalityofIssues

Bodle, Robert (2011). Regimes of sharing. *Information, Communication & Society*, 14(3), 320–337. http://dx.doi.org/10.1080/1369118X.2010.542825

Bollier D (2010) *The Promise and Peril of Big Data*. Washington, DC: The Aspen Institute. Available at: http://www.ilmresource.com/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf

Borra E, Weltevrede E, Ciuccarelli P, et al. (2014) Contropedia – the analysis and visualization of controversies in Wikipedia articles. In: *OPENSYM: proceedings of the 10th international symposium on open collaboration*, Berlin, 27–29 August.

Boullier D (2013.) Le 'hard' du 'soft': la matérialité du réseau des réseaux. *CERISCOPE Puissance*. Available at: http://ceriscope.sciences-po.fr/puissance/content/part2/le-hard-du-soft-la-materialite-du-reseau-des-reseaux

Bounegru L, Gray J, Venturini T, et al. (2018) *A Field Guide to "Fake News" and other Information Disorders*. Available at https://ssrn.com/abstract=3024202

Bowker GC (2013) Data flakes: an afterword to 'raw data' is an oxymoron. In: Gitelman L (ed.) *'Raw Data' Is an Oxymoron*. Cambridge, MA: The MIT Press, pp. 167–173.

Boyd D and Crawford K (2011) Six provocations for big data. Paper to be presented at Oxford Internet Institute's – A Decade in Internet Time: Symposium, Oxford, 21 September. Available at: http://doi.org/10.2139/ssrn.1926431

Boyd D and Crawford K (2012) Critical questions for big data. *Information, Communication & Society* 15(5): 662–679.

Bucher T and Helmond A (2017) The affordances of social media platforms. In: Burgess J, Poell T and Marwick A (eds) *The SAGE Handbook of Social Media*. London: SAGE.

Burrows R and Savage M (2014) After the crisis? Big data and the methodological challenges of empirical sociology. *Big Data & Society*. Epub ahead of print 10 July. DOI: 10.1177/2053951714540280

Cacciatore MA, Scheufele DA and Iyengar S (2016) The end of framing as we know it and the future of media effects. *Mass Communication and Society* 19(1): 7–23.

Carey JW (1967) Harold Adams Innis and Marshall McLuhan. *The Antioch Review* 27(1): 5–39.

Castells M (2009) *Communication Power*. Oxford: Oxford University Press.

Drucker J (2011) Humanities approaches to graphical display. *Digital Humanities Quarterly* 5(1): 1–20.

Flyverbom M and Madsen AK (2015) Sorting data out – unpacking big data value chains and algorithmic knowledge production. In *Die Gesellschaft der Daten: Uber die digitale Transformation der sozialen Ordnung*. Bielefeld: Verlag, pp. 123-144. Available at: http://doi.org/10.4135/9781412985871

Frischmann B (2001) Privatization and commercialization of the Internet infrastructure. *The Columbia Science and Technology Review*. Epub ahead of print 12 November. DOI: 10.2139/ ssrn.255423

Fuller M (2005) *Media Ecologies*. Cambridge, MA: The MIT Press.

Gans HJ (1993) Reopening the black box: toward a limited effects theory. *Journal of Communication* 43(4): 29–35.

Gerbaudo P (2012) *Tweets and the Streets: Social Media and Contemporary Activism*. London: Pluto Books.

Gerlitz C and Rieder B (2018) Tweets are not created equal: Investigating Twitter's client ecosystem. *International Journal of Communication* 12: 528–547.

Gibson JJ (1986) *The Ecological Approach to Visual Perception*. Hove: Psychology Press.

Gillespie T (2010) The politics of 'platforms'. *New Media & Society* 12(3): 347–364.

Gillespie T, Boczkowski PJ and Foot KA (2014) *Media Technologies: Essays on Communication, Materiality and Society*. Cambridge, MA: The MIT Press.

Gray J, Bounegru L, Milan S, et al. (2016) Ways of seeing data: towards a critical literacy for data visualizations as research objects and research devices. In: Kubitschko S and Kaun A (eds) *Innovative Methods in Media and Communication Research*. London: Palgrave Macmillan, pp. 227–251.

Hargittai E and Hsieh YP (2013) Digital inequality. In: Dutton WH (ed.) *Oxford Handbook of Internet Studies*. Oxford: Oxford University Press, pp. 129–150.

Havens T and Lotz AD (2012) *Understanding Media Industries*. Oxford: Oxford University Press. Helmond A (2015) The platformization of the web: making web data platform ready. *Social Media + Society*. Epub ahead of print 30 September. DOI: 10.1177/2056305115603080

Innis HA (1986) *Empire and Communications*. Toronto, ON, Canada: Dundurn Press.

Innis HA (2008) *The Bias of Communication*. Toronto, ON, Canada: University of Toronto Press. Katz E (2001) Lazarsfeld's map of media effects. *International Journal of Public Opinion Research* 13(3): 270–279.

Kooti F, Gummadi KP, Mason WA, et al. (2012) The emergence of conventions in online social networks. In: *Proceedings of the sixth international AAAI conference on weblogs and social media (ICWSM)*, Dublin, Ireland, Spain, 4–7 June. Palo Alto, CA: AAAI Press.

Latour B (1985) Visualisation and cognition: drawing things together. In: Kuklick H (ed.) *Knowledge and Society Studies in the Sociology of Culture past and Present*. Greenwich, CT: Jai Press, pp. 1–40.

Latour B (1998) Thought experiments in social science: from the social contract to virtual society. In: *1st virtual society? Annual public lecture*, Brunel University London, London, 1 April. Latour B (2005) *Reassembling the Social an Introduction to Actor-network Theory*. Oxford: Oxford University Press.

Latour B and Woolgar S (1979) *Laboratory Life: The Construction of Scientific Facts*. Princeton, NJ: Princeton University Press.

Lazer D, Kennedy R, King G, et al. (2014) The parable of Google Flu: traps in big data analysis. *Science* 343(6176): 1203–1205.

Lazer D, Pentland A, Adamic L, et al. (2009) Computational social science. *Science* 323(5915): 721–723.

Leskovec J, Backstrom L, and Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 497–506). Paris, France: ACM. Available at: http://doi.org/10.1145/1557019.1557077

Lury C and Wakeford N (2012) *Inventive Methods: The Happening of the Social*. New York: Routledge.

McCormick TH, Lee H, Cesare N, et al. (2015) Using Twitter for demographic and social science research: tools for data collection and processing. *Sociological Methods & Research* 46: 390–421.

McLuhan M (1964) *Understanding Media: The Extensions of Man*. Cambridge, MA: The MIT Press.

McLuhan M (1967) *The Medium Is the Massage: An Inventory of Effects*. Berkeley, CA: Gingko Press.

Madsen AK (2012) Web-visions as controversy-lenses. *Interdisciplinary Science Reviews* 37(1): 51–68.

Madsen AK (2015) Tracing data -paying attention: interpreting methods through valuation studies and Gibson's theory of perception. In: Kornberger M, Justesen L and Madsen AK (eds) *Making Things Valuable*. Oxford: Oxford University Press, pp. 257–278.

Mandiberg M (ed.) (2012) *The Social Media Reader*. New York: New York University Press. Marcus G and Davis E (2014) Eight (no, nine!) problems with big data. *The New York Times*, 6 April. Available at: http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html

Marres N (2012) The redistribution of methods: on intervention in digital social research, broadly conceived. *The Sociological Review* 60: 139–165.

Marres N (2017) *Digital Sociology: The Reinvention of Social Research*. Cambridge: Polity Press. Marres N and Gerlitz C (2016) Interface methods: renegotiating relations between digital social research, STS and sociology. *The Sociological Review* 64(1): 21–46.

Marres N and Moats D (2015) Mapping Controversies with Social Media: The Case for Symmetry. *Social Media + Society 1*(2): 1–17. Available at: http://doi.org/10.1177/2056305115604176

Marres N and Weltevrede E (2013) Scraping the social? Issues in real-time social research. *Journal of Cultural Economy* 6(3): 313–335.

Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Houghton Mifflin Harcourt.

Monod J (1970) *Le Hasard et la Nécessité*. Paris: Seuil.

Moretti, F. (2013). "Operationalizing": or, the function of measurement in modern literary theory. *New Left Review*, *84* (November-December): 1–13.

Munk AK (2013) Techno-anthropology and the digital natives. In: Botin L and Børsen T (eds) *What Is Techno-Anthropology?* Aalborg: Aalborg University Press, pp. 287–309.

Niederer S (2013) Global warming is not a crisis! Studying climate change skepticism on the web. *Necsus*, 3 June. Available at: http://necsus-ejms.org/global-warming-is-not-a-crisis-studying-climate-change-skepticism-on-the-web

O'Neil C and Schutt R (2013) *Doing Data Science*. Sebastopol, CA: O'Reilly Media.

Paßmann J and Gerlitz C (2014) Good' platform-political reasons for ‚bad' platform-data. Zur sozio-technischen Geschichte der Plattformaktivitäten. *Mediale Kontrolle unter Beobachtung* 3(1): 1–40.

Peters JD (2015) *The Marvelous Clouds: Toward a Philosophy of Elemental Media*. Chicago, IL: The University of Chicago Press.

Rein K and Venturini T (2018) Ploughing digital landscapes: How Facebook influences the evolution of live video streaming. *New Media & Society, Forth*. Available at: http://doi.org/10.1177/1461444817748954

Rieder B (2013) Studying Facebook via data extraction: the Netvizz application. In: *Proceedings of the 5th annual ACM web science conference*, Paris, 2–4 May.

Rieder B and Röhle T (2012) Digital methods: five challenges. In: Berry DM (ed.) *Understanding Digital Humanities*. Basingstoke: Palgrave Macmillan, pp. 67–84.

Rogers R (1999) *Technological Landscapes*. London: Royal College of Art.

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

Rogers R (2013) *Digital Methods*. Cambridge, MA: The MIT Press.

Rogers R (2017) Foundations of digital methods: query design. In: Schäfer MT and van Es K (eds) *The Datafied Society: Studying Culture through Data*. Amsterdam: Amsterdam University Press, pp. 75–94.

Rogers R and Marres N (2002) French scandals on the web, and on the streets: a small experiment in stretching the limits of reported reality. *Asian Journal of Social Science* 30: 339–353.

Ruppert E, Law J and Savage M (2013) Reassembling social science methods: the challenge of digital devices. *Theory, Culture & Society* 30(4): 22–46.

Sanchez-Querubin N, Schuhmacher J, Traldi G, et al. (2016) #Recovery: recovering in and with Tumblr. *Digital Methods Summer School*. Available at: https://wiki.digitalmethods.net/Dmi/SummerSchool2016RecoverySelfie

Savage M and Burrows R (2007) The coming crisis of empirical sociology. *Sociology* 41(5): 885–899.

Shifman L (2013) *Memes in Digital Culture*. Cambridge Mass: MIT Press.

Sloan L, Morgan J, Burnap P, et al. (2015) Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PLoS ONE* 10(3): e0115545.

Srnicek N (2017) *Platform Capitalism*. Cambridge: Polity Press.

Staiger J (2005) *Media Reception studies*. New York: New York University Press.

Venturini T, Jacomy M, Meunier A, et al. (2017) An unexpected journey: a few lessons from sciences Po médialab's experience. *Big Data & Society*. Epub ahead of print 11 August. DOI: 10.1177/2053951717720949

Venturini T, Jensen P and Latour B (2015) Fill in the gap: a new alliance for social and natural sciences. *Journal of Artificial Societies and Social Simulation* 18(2): 11.

Venturini T, Laffite NB, Cointet J-P, et al. (2014) Three maps and three misunderstandings: a digital mapping of climate diplomacy. *Big Data & Society*. Epub ahead print of 5 August. DOI: 10.1177/2053951714543804

Venturini T, Ricci D, Mauri M, et al. (2015) Designing controversies and their publics. *Design Issues* 31(3): 74–87.

Weltevrede E (2016) *Repurposing digital methods: the research affordances of platforms and engines*. PhD Thesis, University of Amsterdam, Amsterdam. Available at: https://wiki.digitalmethods.net/pub/Dmi/RepurposingDigitalMethods/Weltevrede_RepurposingDigitalMethods.pdf

Weltevrede E and Borra E (2016) Platform affordances and data practices: the value of dispute on Wikipedia. *Big Data & Society* 3(1): 75–94.