

2 Big data challenge for social sciences and market research

From society and opinion to replications

Dominique Boullier

Translated by Jim O'Hagan

A new generation of social sciences is knocking at the door. To put it simply, marketing and computer sciences take ownership and generate monitoring of the tools for social life, in the form of monitoring brands, reputations, communities, social networks, etc., which can do without the interpretations and models of the social sciences because they compensate with computing power and the unprecedented traceability of “big data”. Their main concern remains the action and reaction, not the analysis or understanding, as the traditions of sociology and other social sciences have defined. Traces rather than data, reactivity rather than reflexivity, the digital world finds itself shaped by principles that leave less and less room for social sciences.

First, we will discuss the peculiar status of these traces to understand the justifications and limitations of such hype. Then we will compare our times of tremendous technical change to those of the 1930s, when sampling was invented and opinions made visible in the same socio-technical and marketing moves. Finally, another time frame will serve for comparison, when “society” began to be thought of as an entity of its own and computed with novel devices at the same time. Finally, we will discuss how social sciences and market research can repurpose these digital sources for their own goals, while brands use them for their reactivity requirements. The three ages of social sciences are not mere questions for quantification sociologists: they frame our collective reflexivity. The table of their features gives striking evidence of the correspondence between these three ages.

The digital age

Neither people nor identities, traces are the raw material

For many years, computer science (in a way extended with social networks) has calculated and modelled society as if the traces collected allowed access to the “truth” about individuals in a more effective way than polls, surveys and censuses. Consider two examples, one academic and the other commercial:

- “The Web does not just connect machines, it connects people” (Knight Foundation, 14 September 2008). This was declared by Sir Tim Berners-Lee, founder of the World Wide Web (WWW) in 1991 with René Caillau, emphasizing the transition to a dimension of networks that are neither technological (Internet) nor documentary (WWW), but rather social (GGG, for Global Giant Graph).
- Facebook, for its part, has managed the tour de force of “normalizing” in terms of the actors themselves and the declaration of their true identity. That is to say, those provided by the civil state, their name and surname, in opposition to the tradition of anonymity on the Web. The platform thus claims to become the world of reference or even a civil-status alternative, competing with Google in this regard.

However, there is no guarantee whatsoever of any connection between identities on Facebook or Berners-Lee’s “people” and persons identified by their civil registry. What are connected are merely the retrieved accounts and data, and these are only the traces of activity from an entity that could possibly take on the form of civil status. This uncertainty should lead to a careful examination of results in order to differentiate registered accounts and active or engaging accounts, and prevent anyone from allegations about “society” or “people” when using social networks analytics. For the scores that classify sites on a search engine such as Google, the resulting topology of sites and blogs never discusses their contents as such, but the inbound and outbound links that produce a rank of authority or hub, as defined in the network topology (Kleinberg et al., 1998) and not a civil status.

What we mean by “traces” should be noted here, in order to distinguish them from data. Traces can range from signals (“raw” generated by objects) to unstructured verbatim. Traces can be exploited in databases (links, clicks, likes, cookies)¹ and by operators or platforms, but they can also be captured independently of this through the application programming interface (API) and, as such, falling outside relational databases. Traces are not necessarily pre-formatted for a specific calculation, nor are they dependent upon aggregation that can then be applied. It is easy to argue that, despite everything, “behind” these sites or “behind” these clicks there are most certainly people, but that does not alter the fact that the algorithms do not take this into consideration and that, furthermore, no guarantees can be given in this regard. Traces understood in the restricted sense are produced by platforms and digital-technological systems, but are not “signs” or evidence of anything other than themselves as long as relationships with other attributes are not created and validated. This differs radically from the data that can be recovered en masse from client files or administrative acts. Certainly, the big data methods for calculating can be applied here in both cases, but the traces are *a priori*, independent of other attributes, in particular socio-demographic factors that are rarely mobilized in correlations sought between traces. Relationships with more conventional parameters in data sciences are limited to

22 *Dominique Boullier*

time (a timestamp) and location (geolocation tags), which allow for the production of timelines and maps that become simplified modes of representation for traces.

Traces are produced by platforms

Amazon and Apple do not focus on the same features as Facebook or Google (since the Web is no longer distributed but rather monopolized by these four GAFA platforms that centralize the majority of traffic, with Twitter extending the traces industry). It is not people who are put into relation with each other, but, above all, tastes (books or music originally), expressed by traces of purchases and choices, which can be treated en masse to produce patterns and profiles, independent of personal information. It should certainly not be forgotten that all these platforms, without exception, are also very fond of civil status-type data, phone numbers and other highly attractive resources to advertisers to whom they are resold. The ensuing marketing methods are largely based on the addressing of mass advertising or email to IP addresses or email accounts that have clicked on an article (retargeting), but much more rarely via sophisticated links with other attributes of the supposed people attached to these addresses or those clicks (profiling).

Traces of digital behaviour are thus a particularly profitable “raw material”, without the need to appeal to the social sciences. How should the social sciences deal with this situation? Two options are open to them: either they are confined to their world of administrative data, surveys and polls, relativizing the interest of such traces and focusing on data; or they decide to take the bull by the horns and take these traces as raw material provided for “repurposing”, as proposed by Richard Rogers (2013). Thus, they must accept being dependant on the platforms that produce these traces, without being able to hold any weight over their formatting, or even being totally dependent on the conditions of production for such data, which may vary over time and across platforms. The powerful viral phenomenon specific to the Facebook platform and its “likes” mechanism nevertheless does not leave researchers indifferent, since they are so spectacular, such as when pages generate substantial likes in a few days. This arouses all types of analyses from the most constructivist and critical (“all the likes are bought”, i.e. artefacts), to the most realistic (“it is solid proof that opinion, and even ‘the people’ think that way”). The limited quality of the traces is observable on all platforms, but these limits may be intrinsic when they do not meet the criteria for traceability that we consider crucial in order to exploit them, or extrinsic when we criticize their lack of reliable relation to the “real” world. The latter stance is found in boyd and Crawford in relation to Twitter:

Some users have multiple accounts. Some accounts are used by multiple people. Some people never establish an account, and simply access Twitter via the web. Some accounts are “bots” that produce automated

Big data in social sciences and market research 23

content without involving a person. Furthermore, the notion of an “active” account is problematic. While some users post content frequently through Twitter, others participate as “listeners”. Twitter Inc. has revealed that 40 percent of active users sign in just to listen.

(Boyd and Crawford, 2011)

Other studies (Driscoll and Walker, 2014) tested the data produced from various access methods offered by, for example, Twitter, and showed that the Search API, the Streaming API, and the Gnip PowerTrack (paid service) provide very different results, the latter method collecting a much larger number of tweets, but not uniformly according to the requests! This means that the traces collected are entirely dependent on the mechanisms of collection. This is not surprising, although we do tend to forget this for other, older methods that have become conventional.

The brands’ grip on traces

Where does this fascination with traces – despite their limitations – come from, compared with data from registries and surveys? The traces are actually a key resource for brands to monitor the effects of their actions on the public. Reputation and notoriety no longer translate audience measurement, which would be a simplistic import of measures, largely built for the mass media. On networks, one must measure both a form of audience (the reach), the most basic activities of its uncertain public (likes, stars), and also more sophisticated activities such as comments, which constitute what is called “the engagement rate”. Brands are fond of these traces and they fuel the turnover of all these platforms, and thus, of the entire Web. The opinion-mining and sentiment-analysis tools (Boullier and Lohard, 2012) are thus the answer to the marketer’s anxiety after the product launch. However, the extension of this brand domain reaches all activities, whether commercial, cultural, political, institutional or even interpersonal, when each must measure excellence with rankings, as researchers are requested to do (Bruno and Didier, 2013). Thus, it is the brands’ methods that take precedence everywhere and impose their law and their pace, even on public services.

However, what primarily concerns these brands is not structured and constructed data to test, for example, causality, but many traces that function as indicators and alerts, even approximate, not at the individual level but at the level of trends. Similarly, it is not reflexivity that is sought but primarily reactivity, the ability to determine which lever to act upon in relation to the dimensions (features) of the brand that are affected. The closer the relationship with the devices that monitor the activity and the offers made to the customer, such as in customer relationship management (CRM) systems, the more efficient the reactivity. Algorithms that decide the prices of advertisement placement no longer depend on negotiations or decisions made by experts in pricing and market segmentation, but instead on the previous amount of

24 *Dominique Boullier*

traces collected by the systems (clicks, for instance) and assembled in correlations that generate automated decisions. These methods and the calculation devices that have been built for market purposes were imported from the financial market, where reactivity is the key factor, to the point of high-frequency trading, where expectations and moves on the markets can be manipulated automatically at a millisecond pace. The political world is now caught up in the spiral of reactivity and its addiction to tweets leads us to consider that we have entered the era of *high-frequency politics* (Boullier, 2013).

We have drawn up a table that merits systemization. Digital networking generates:

- Traces
- Assembled and formatted by platforms
- For brands
- With a view to reactivity
- In order to produce rankings or patterns.

This situation is not new. Two other key moments in the existence of the social sciences, especially sociology, market studies and political science, must be paralleled on the same basis to understand the scope of the changes underway. These two main periods of quantification for societies (Desrosières, 1998) must teach us how new methods and principles can be arranged in such a way that they transform themselves “into socio-technical conventions”. The emergence of big data can be as challenging for social sciences as were sampling methods in the 1930s, for instance. Big data’s ability to become a shared method to quantify social phenomena may produce the same lasting effects.

The construction of “opinion”²

The contemporary situation is undoubtedly not that far from a key moment in the history of the social sciences, especially when dealing with consumption behaviour issues that would help us understand what is happening: this is why we will review some features of the period to guess the equivalent in our times. If we gave the current era of digital traces the label “3G” for “third generation”, we would then have to give the emergence of public opinion in the late 1930s the label “2G”. Indeed, in 1936, George Gallup was able to predict the election of Franklin D. Roosevelt over Alf Landon with a study of 50,000 people. Thus, this dramatic gesture founded the reliability of the survey and of investigative sampling methods,³ which certainly sacrificed the exhaustiveness of inquiries on entire populations but managed to produce correct results provided that the terms of *representativeness* were respected. The issues of sampling were addressed previously by the social survey of Rowntree (1901) studying urban poverty in York and using Booth’s “poverty

line” as a statistical marker. However, as Converse (1987) told the story, it is certainly in the context of the mass media that their importance was recognized and their method systematized. With Ogilvy, Gallup studied film audiences and then with Crossley at Young & Rubicam, he studied radio audiences using telephone interviews before even making a proposal to conduct the election polls (Converse, 1987). Considerable media transformation, and the mass media (radio at the time), has established the conditions for the emergence and validation of a survey technique, which thus opens up a whole new era for marketing and political science. Moreover, it is “public opinion” that takes on a measurable existence with these sampling methods.

Communication agencies such as polling organizations indeed cannot live solely from their campaign activities, even if these do bring them high visibility and notoriety. From the outset, their target is constituted by the mass media for one essential reason: audience measurement becomes key to the distribution of advertising space. This has been true since the dawn of radio and then later with television (in 1941 the first ads were aired on American television for Bulova watches, during a baseball game). However, these measures also allow us to monitor the effects of these campaigns on the minds of consumers, giving an unprecedented boost to marketing that drives increasingly sophisticated communication strategies (Cochoy, 1999). Agencies that provide the main reliable feedback on audiences are used to design the programmes and the ads as well, targeted at the same populations and generating revenue from companies. Brands are thus present from the beginning in methods of inquiry into opinion via sampling from the moment when such investigations were aimed primarily at mass-media audiences.

Public opinion exists. I measured it!

The work done by Gallup for the operational side, and Lazarsfeld for the scientific side, is therefore not a simple marketing operation or a facelift for the social sciences: it provides whole societies with methods with which to auto-analyse and to represent themselves as opinions. Tarde (1989 [1901]) certainly highlighted the importance of these views; it is only when the metrics are established and produced in a conventional way that opinion finally exists. Only the media’s control and their ability to produce a unified public in a national territory enabled this methodological assembly to hold on. The “whole” referred to by the polls is in fact originally the *public* formed by the media, which allows the audience to emerge as *public opinion*, and to make it permanently visible and measurable with the aim of being exploitable for brands to measure the influence of their campaigns. The parts (Latour et al., 2012) that are individual expressions are pre-formatted to be recordable and calculable, but the link between the parts and everything else is made only by the pollsters’ black boxes. The rigorous, scientific precautions are upheld through “confidence intervals” (defined by Neyman in 1934), which keep a reference on the comprehensiveness of the studied population. Such

26 *Dominique Boullier*

successful convention work focuses on the same assemblages of mediations already mentioned for traces:

- The “surveys” and “polls” (from individual expressions framed by questions and thus made calculable)
- Assembled and formatted by pollsters
- Guarantees the representativeness of samples (sampling)
- Made for the media
- Made for the purpose of monitoring
- Generates public opinion (and audiences).

The fabrication of “society”

This historical reference to opinion might seem too close to the digitally networked world because of the involvement of the media and brands. Therefore, the world of traces produced on the Web may ultimately be restricted to a permanent extension of the domain of brands and other metrics. Yet, to us it seems that another historic moment for the social sciences would allow us to complicate the panorama and perceive it in the long term. Let us effectively pretend here that Durkheim succeeded in an operation identical to when Gallup and Lazarsfeld invented “public opinion” because he managed to make “society” exist. If the conventional nature of the concept of opinion may still be admitted, it stands that evidence about society does not bear discussion, not only for academia but also for the layman and the experts of markets. This is especially true since the term did not begin with Durkheim, although its history is not so long. Durkheim’s early work on the “division of labour in society” (1893) was not based on statistical methods, but instead laid the foundation for a model of social types, aggregated in mechanical and organic solidarity. With *Suicide* (Durkheim, 1897), the method was set up to extend the discussion of the types that would reveal anomia to be a problematic situation. However, reliance on data records produced by states, from their various components (ministries, prefectures, governments), becomes key to the demonstration. It is these aggregates that are explained or explanatory, using a method of comparison between countries, regions, counties or districts where possible and necessary. The method depends entirely on the available data and cannot afford to criticize or question the procedures for the production of these data, despite the countless limitations identified upon publication. By organizing all his systems of proof around these national administrative statistics, Durkheim finds a quantitative analogue for his conceptual choice that puts “society” in a separate status from all manifestations and individual behaviours. Durkheim’s *whole* becomes an entity of the second degree, “society” (Latour, 2005), while the censuses and other state-data registers simply conduct the task of recovering individual, administrative events (marital status, judicial

Big data in social sciences and market research 27

procedures, etc.), formatted in identical categories and aggregated to reveal the behaviour of populations. All Durkheim's force of conviction would have been to make these statistics exist as equivalent to his "society".

The statistical apparatus makes society visible in the same way that the survey makes opinion visible; regardless of the statistical validity, the framing that operates here gains power. It is indeed necessary to notice that a form of "objective alliance" was formed between data producers from the state administrations and the emerging social sciences. Together they produced the entity "society" as the object to be tracked by the state for the government and to be explained for scientific reasons. The result is the widely shared and obvious fact that "society" exists and the methods that allow it to do so have no grounds to be questioned because they demonstrate both their scientific and operational value; they are "tools of proof" and "tools of government", as Desrosières (2014) put it. These processes and alliances look absolutely identical to those we encounter between the media and the polling organizations which get on well with one another in order to fabricate opinion and make it seem natural, *taken for granted*, after a long work of implementing conventions.

Technical devices should be considered as parts of the assemblage for these conventions. In 1890, Hollerith used his machine to conduct the American census because the Census Bureau had failed to finish processing the previous 1880 census before starting the next. Hollerith's company would later be transformed into IBM by Watson in 1926 and would spread among all countries' administrations. This specific feature of new devices enrolled in this new quantification era is exactly what is at stake in the digital revolution, as it was at the times of phone lines and mass media for opinion. This socio-technical setting is well known in science and technology studies, and becomes very useful in times of ongoing innovations that generate so much disorientation. The ability of socio-technical designs to assemble a specific set of features, to make them last long after the innovation breakthrough, and to become a "taken-for-granted" part of the "social environment" are what make technologies powerful in maintaining "a sense of the social structure", as Schütz (1962) used to say.

Durkheim's performance would have been to hold an assemblage of very powerful mediations:

- Censuses
- Assembled and formatted by public administrations
- Under guarantee of exhaustiveness
- For states
- With a government in mind
- To produce "society" (based on population)
- Using tabulating calculation machines.

28 *Dominique Boullier*

What the social sciences can do with the digital, and what the digital does to the social sciences

Replacing digital transformations in the long history of the social sciences allows us to better understand contemporary movements in the use of traces. Three positions are possible:

- One that aims to take up the course of the social sciences of previous generations and applies their methods and concepts of “society” and “opinion” to trace the Web.
- One that accepts this new world of traces, immersing itself in its demands and principles by abandoning traditions and scientific requirements.
- One that confronts the radical novelty of this socio-technical configuration and attempts to understand what the social sciences’ place might be in the production of new conventions to exploit these traces.

The digitalization of opinion and society traditions

The first direction consists of taking up well-known methods and concepts and applying them to traces collected on the Web, or, even more sensibly, exploiting the potential of digital networks to implement exactly the same methods. Thus, surveys via questionnaire and polls are not only computer-aided on platforms equipped for this, but can be conducted entirely on the Internet and sometimes permit the recovery of enough of the respondents’ socio-demographic attributes to ensure the sample’s representativeness. These online surveys are ideal for making opinion more responsive and tracked on a more frequent basis. Thus, the digital amplifies (Eisenstein, 1983) “the reality of public opinion”. Sampling is used in online surveys for market research that makes the rationale of consumers tastes, opinions and judgements appear through their formatted individual expressions. Likewise, input modes for censuses could be equipped with computer terminals to speed up and standardize the collection of data, which makes the now visible “society” even more reliable. However, Web studies, a by-product of the social sciences, implement the same framework on these new media formats: economic studies, preferably from Google queries, studies of sociability, longitudinal monitoring of “communities” around themes or specific sites, “opinion mining” and “sentiment analysis” methods (Boullier and Lohard, 2012) are implemented to increase the monitoring of public opinion or the identification of trends in consumer behaviours. In this approach, digital traces on social networks or blogs are just one more way to get access to these opinions that are not questioned and not dependant on the platforms where they appear.

The end of social theory?

Another orientation is available, a rather radical one when phrased by Chris Anderson. The editor of *Wired* magazine raised concerns with his 2008 short paper entitled “The End of Theory”. Extending his provocative statement to social networks analysis, for instance, one can say that Facebook likes do not need theory. The platform picks up traces of the actions and the clicks of Internet users or machines, in a standardized format. It then aggregates them and produces a score that is displayed and can be used by the platform to show trends that guide the placements of advertisers who also seek to achieve certain effects and optimize their investment or communication choices. In a simplified format, this is the string of events that is produced. Social theory has virtually no use in such an operational system, where the performative mechanism works in almost the same way as audience measurement. Some have tried to develop a critique showing that the likes aggregate very different sorts of behaviour, including even purchased likes, but this hardly concerns operators, platforms or advertisers. Their action/reaction works in the performative mode, where the likes unearth a reality that will initiate strategies to influence the likes, in a self-referential cycle to which one could also assign audience ratings. However, in the case of audience ratings, all advertisers and programmers have agreed on stable criteria, producing a shared agreement. Evidence of this has come to forcefully impose itself every morning in the direction of programmes in the mass media. The platforms of social networks and advertisers have not yet reached a stable compromise, which explains the proliferation of services that claim to be standard (such as Klout) and want to become the Nielsen of these measures. However, no theory is necessary for all these institutes if there is high statistical quality (which is seldom the case!), as any theoretical advance in audiences and their processes would require a renegotiation of the agreements, which remains the only validity criterion of any theory. It is easy to see the difference between these principles and the traditions of the social sciences, as G. Bowker does to show their extreme reductionism:

If I am defined by my clicks and purchases and so forth, I get represented largely as a person with no qualities other than “consumer with tastes”. However, creating a system that locks me into my tastes reduces me significantly. Individuals are not stable categories – things and people are not identical with themselves over time [...] The unexamined term the “individual” is what structures the database and significantly excludes temporality.

(Bowker, 2014, p. 1797)

Bowker has cause for concern from the point of view of “society”, but the third generation of the social sciences is not so much interested in “society” as in other processes created by other devices, but which, nonetheless, cause us

30 *Dominique Boullier*

to act. Brands, reputations and recommendations, as exploited by Amazon, can certainly be forcefully re-injected into a “society” matrix to make them say what they are not made for saying, but they also say something of themselves, from another world: that of the power of recommendations and contagions that the social sciences are reluctant to understand. Censuses are no longer the vital statistics that are the reference base for the third generation of the social sciences, but are instead totally agnostic traces about the entities that are “behind” because all act almost equivalently and cause the others to act.

The properties of the third generation of the social sciences

What the third generation of the social sciences could be remains to be observed or even imagined if they:

- assume the radically new character of these heterogeneous traces without falling back on the status of traces or symptoms of something “truly” social (“society” or “opinion”; “market” being a hybrid of both);
- do not get caught up in the self-referential production system/monitoring of traces that dispenses with theory because of other aims.

We adopt a radical empiricism approach (James, 1890) in following the digital traces at their face value, looking for what they are (traces produced by platforms), and how they transform and are transformed by the very milieu they live within, refusing to reduce them to an equivalent of any other phenomena in society or in opinion. This constraint strongly limits the power of “explanation” of these traces, but complies with other approaches that consider “platforms bias” as a component of the analysis. These high-frequency propagation phenomena cannot separate traces (elements) and platforms (milieux), and must account for the distribution of agency, which is never settled as an *a priori*. Following them allows us to compute these imitation processes, which, as Tarde said, include opposition and invention (Tarde, 2001 [1895]). This approach does not contest the legitimacy of other analyses of the social as long-term social structures, or as opinion movements of mid-range frequency (fashion and elections have almost the same wavelength), and should help more traditional social sciences seriously consider the extension of the social to new entities, these *replications* that could not be captured before the digital era. We would like to avoid two traps at the same time: using digital traces for the documentation of the “social-as-usual” (society or opinion); or reducing them to a set of clever tricks of methods. Our responsibility for the social sciences to be able to play their full role in the digital world dominated by platforms and brands is to build the conventions for a new layer of social sciences maintaining the requirements of scientific reflexivity.

The affinity of big data’s quality criteria with the requirements of the social sciences is quite striking. They are often summarized with the 3Vs: volume, variety and velocity.

Volume and exhaustiveness

The volume, in some way, mimics the need for exhaustiveness famous in social sciences. However, this results in a somewhat limited mode, because nobody and nothing can define the boundaries of the universe of data collected. We clearly need to mourn the death of exhaustiveness when using Web traces, but that does not mean dispensing with the laying down of all conventional frameworks for social science's approach when dealing with digital traces.

Variety and representativeness

The second criterion, variety, is also a form of transcription for the representativeness requirements that allow all social sciences to proceed with inquiries and surveys based on sampling. Again, the test is a loose version of representativeness, which assumes that we accept a *sufficient* level of variety. The establishment of a set of sources (sourcing) in studies of the Web should then adhere to some criteria, specific to digital methods and to the field of study. Our work on opinion mining has led us to consider that no description of social society, social opinion or social traces can be produced “in general” on digital networks. The social sciences must agree to deal solely with “issues” (Marres and Weltevrede, 2013), on the focal points of attention, or on “oriented and situated engagements” (Hannerz, 1983), for which traces that are specific to each outcome or each engagement can be kept digitally.

Velocity and traceability

The last criterion, velocity, hardly finds a parallel in the first and second generations of the social sciences. Indeed, these dynamic processes were neither their *forte* nor their concern. It was essential to seek primarily to represent the positions at a given moment *t*, to show the strength of “society” on the diversity of individual behaviour, or to show how public opinion and consumer appraisal are structured beyond singular expressions obtained in surveys. Certainly, through a longitudinal follow-up of the same populations or by reusing the same questionnaire, it is possible to deliver the equivalent of dynamism, but without ever being able to track the mediations that would produce these changes. Velocity seems outside the scope of conventional approaches.

However, a branch of Web science has also seized upon the issue of velocity in its own way by exploiting the *meme* traces that spread on the Web. It is very significant that Kleinberg, the very man who exported scientometric methods to the study of Web topology – methods that were taken up by Google – has been interested in the technical development of a “meme tracker” with Leskovec since 2002 (Leskovec et al., 2009). Their most famous study looked at the propagation of citations throughout Barack Obama's

32 *Dominique Boullier*

presidential campaign, which allowed them to achieve a spectacular visualization of the focus of attention in rapid mounting and descending curves (*streams and cascades*) around certain incidents during the campaign. Their method aggregated all types of traces that these citations leave, and treated them as strings of characters that can be found throughout the entire Web. This produced a metric anchored in time, day to day and even minute to minute with Twitter. (The unit of measurement has become the tweet-per-second.) Taking memes into consideration seems promising to us, provided that we also track the transformations-translations of these memes (derived from memetics, Dawkins, 1976; Blackmore, 1999) in different environments.

Therefore, it becomes possible to find an equivalent for the velocity of big data: *traceability*. This becomes the essential quality criterion for entities that can be studied.

The third generation of the social sciences will hardly be able to do anything but associate with digital platforms and brands to produce a science of traces which are then treated as “*replications*”, as we propose below. The traces produced are platform-dependent; we can hardly expect to modify them at the source. Nevertheless, it is possible to exploit the traces produced by the platforms by diverting them from the purpose for which they were designed. The rule here is that we do not take any explanation at another level or another world into account, but that we might compare propagation speeds, rhythms and possible transformations (e.g. contamination of other areas, etc.). The difference should be the ability to see the processes that have not yet been identified, either because of the limits of pre-digital technology or the targets adopted by previous generations of social sciences. As R. Rogers has pointed to in his pioneering work, this “repurposing” of traces (Rogers, 2013) will suppose a fine-tuning of the “query design”, on Google or on any API. For Rogers, this should rely on a well-defined hypothesis and not only follow the opportunity of inference offered by big data and machine-learning technologies. N. Marres seems more concerned by the critics about the overwhelming dependency of scholars on the platforms that deliver the data. The way she handles socio-political controversies as “issues” has proven very inspiring by designing limits of validity for empirical research. (For instance, no general tracing of tweets or weibos⁴ or any other traces without delineating the arenas made by “issues”.)

From traces to replications⁵

As we have highlighted, the production of traces is directly dependent on platforms that generate their own analyses. While cooperation with these platforms is required in order to reach conventional methods, it is still critical to produce the theoretical framework that would account for this phenomenon of traces. Let us then talk about “*replications*”. It may help social sciences to accept the shift achieved vis-à-vis the notions of actors, strategies and representations. All of these notions have their legitimacy in the context of

other social sciences but do not allow for these circulating entities' *power to act* (the *replications*' agency) to be taken into account. We cannot say *a priori* what the size or status of these entities are, because it is only the mass corpus investigations that allow us to identify them when their vibration emerges from the sensors we exploit, certainly from platforms but also according to *our* objectives.

The principle of a sociology of *replications* relies on the need to follow the elements in order to detect waves, without knowing how they will join together to make a "whole" of variable geometry. The vibration approach allows us to build an infinite combinatory, following extensions, propagations and repetitions, provided we remain focused on "issues" that carry *replications*. It is then necessary to focus on the moments of emergence and not on the peaks that function as aggregates. The object of this science of *replications* is surely the agency of *replications* that spread and end up enveloping us. This is because people are actually traversed by ideas, and ideas do not make us act in the inverse, as Tarde (2001 [1895]) clearly indicated. "The imitation rays first and then the beings, whose existence we infer from the transformation they undergo with the flow of imitation" (Latour, 2011). It is then possible to study the properties of these *replications*, to potentially compare their chances of survival or contamination. This is made possible by differences in their properties, which are always directly related to the "issues" that they carry with them.

We started this work in 1987 with the monitoring of TV conversations and their transposition onto workplaces to make "local public opinion". In two different research projects, we proceeded to monitor the attributes of a photo from the Flickr database in the same way and to track the propagation of cultural signs (songs, flags, landscapes, etc.) in a corpus of websites linked to a region. In the first case, the attributes of the photo (e.g. crossed arms) became tag attractors and thus connected accounts or photos that would have no chance to get connected according to the traditional criteria of social explanatory variables. Tags or icons are *replications* that can be followed, even if they have neither the explicit character verbatim or expressions as in the meme tracker, nor their massiveness.

Potentially, all the traces that we have identified (such as likes, tweets, recommendations, etc.) may be the object of monitoring; however, they require specific tracking tools that exist largely for Twitter only. A detailed review of these tools should be done to ensure that they meet the specifications of a traceability of replications (not just traces for the sake of it or for the reactivity of brands).

Certain approaches from these digitized corpora (not native to the digital) can give an idea of the potential of such methods. Work done on the *n*-grams studied from Google Books (Michel et al., 2011) showed the evolution of the English language (the preterite of irregular verbs). Lev Manovich (2012) created a base of over 1 million examples of manga to compare their most basic attributes, such as the contrast, and produced a unique insight into influences

34 *Dominique Boullier*

between trends. He used similar tools to conduct cultural comparisons between countries from millions of photos on Instagram or from the Maidan square in Kiev. The story of “JeSuisCharlie” as a hashtag and logo all over the social networks and media should be considered as a demonstration of the agency of an entity that is not related to a strategy, an intention, and not accounted for by simple society causations. By focusing on the entities that propagate, we change the distribution of agency and pretend to account for the specific role played by the very features of these messages in the waving of the network.

Some propagation patterns can be detected along with the social features of the nodes (e.g. popular Twitter accounts do not retweet but are retweeted), but the program to detect the agency of semiotic features of the messages (e.g. the tweet or the hashtag) still has to be built. This would give the opportunity of extending the agency apart from structures (the “society”) and individuals (“opinion” extracted from the mind of individuals) to actants, such as messages that frame some issues. This is why a theory of *replications* (or vibrations) follows some principles of actor-network theory (ANT) (Callon and Latour, 1981). This is done by exploring this distribution of agency, thanks to the traceability of these elementary parts that produce the network and perform it, and not just follow social paths already well known by social scientists. The roles played by the founding fathers of ANT (Callon, Latour and Law) can be considered as twofold. The first is quasi-technical through scientometrics, because by describing how scientific facts are produced in this web of citations they paved the way for the topological principles of the Web, as previously mentioned, and also for the whole set of methods of traceability. The second role is of a more philosophical kind, because by “following the actants”, they give the opportunity to repopulate the description of how social worlds are produced and maintained (including non-humans and messages, as well as clicks, likes and so on), and how to escape from the strategist view of the network. When computer scientists and market researchers enter the field of social networks, they immediately look for the “influential” (Rogers, 1983 [1963]) and treat these nodes as strategist actors. This is an extended version of decision-making theories (including rational choice theories and game theory), but it misses the point when trying to account for virality and propagation patterns. The network is always “enacted” and reconfigured along with each issue and with opportunities. This is why ANT always remains on the side of emergence theories, in order to account for what is not a mere effect of structure or of rational choices. Tracing the network is only possible by following the actants, be they artefacts and messages, provided that the observer adopts the right techniques to account for their specific agency.

The consequences of the way market research handles these traces can be listed as follows:

- The process under scrutiny or the central issues are not the same: from segmentation of the market (MR1G, marketing research first generation)

or from trends (MR2G), we move to reputation, that is the way brands use digital traces, or to replications (MR3G) from the market research point of view.

- The entities that are populating this world and being computed are no longer socio-demographical categories nor market segments (MR1G), nor socio-styles, word of mouth, opinion leaders, the influential, consumer journeys, as in the MR2G approach. MR3G focuses on brands and communities when framed by companies and brands, or on viral content features when framed by social sciences. This is an important move, because the agency is not distributed to the same kind of entities. It means that segments or the influential no longer make sense from this point of view. Another set of entities emerge as candidates that had fallen into oblivion before the digital networks shed light on them. This is something content designers are eager to buy, even though they will get back to the omnipotent view of this viral approach, something Jenkins et al. (2013) criticized in favour of their “spreadability”, with some reason.
- The data collection devices are transformed from the CRM, where all targets can be assembled and traced (MR1G), and from polls and focus groups (MR2G), to social listening platforms and community management processes for the brands, and to some kind of meme tracker for the social sciences (MR3G).
- The methods are intrinsically different from “targeting”, borrowing on ballistics (MR1G) and from the influence processes (MR2G), because we focus on propagation patterns for brands as well as for social sciences (MR3G).
- The wavelengths that are investigated are different and do not compete with each other: the long waves of structural social features (MR1G), the mid-length cycles of market trends (MR2G), and the high-frequency waves of memes propagation (MR3G), typical of an emergence approach.

Any market research strategy should be aware of the strategy it adopts among these three generations and should not believe that one choice accounts for all the richness of social life, nor that strategies should be easily combined. We still need to build the conventions to make replications analysis as reliable as censuses and opinion polls. Table 2.1 summarizes these features more clearly.

Conclusion

A table summarizing the three ages of the social sciences allows the consistency of this approach to be made visible, and yet at the same time demands simplification and the elimination of the specificities of each age (see Table 2.2). Remember, however, that we are not dealing with the so-called “qualitative” aspects of the methods of social sciences.

<i>Table 2.1</i> Market research generations				
	<i>First generation</i>	<i>Second generation</i>	<i>Third generation (from brands)</i>	<i>Third generation (from social sciences)</i>
Issues	Segmentation	Trends	Reputation	Replications
Entities	Socio-demographic categories + market segments (e.g. dual income, no kids – DINK)	Socio-styles + word of mouth + opinion leaders + the influential + consumer journey	Brands + communities	Memes + viral content features
Collection devices	CRM	Polls + focus groups	Social listening + community management	Meme tracker
Methods	Targeting (ballistics)	Influence	Propagation	Propagation
Wavelengths	Long waves (structure)	Cycles (market)	High-frequency waves (emergence)	High-frequency waves (emergence)

Table 2.2 The three generations of the social sciences

	<i>First generation</i>	<i>Second generation</i>	<i>Third generation</i>
Concept of the social	Society/ies	Opinion(s)	Replication(s)
Collection devices	Censuses	Surveys, polls	Platforms, big data
Validation principle	Exhaustiveness	Representativeness	Traceability
Co-construction institutions/ research	Registers, inquiries	Audience, polls	Traces, repurposed digital methods
Major players of reference (and financiers)	States	Mass media	Brands
Operational actors	National institutes	Polling organizations	Web platforms (GAFA)
Founding authors	Durkheim	Gallup, Lazarsfeld	Callon, Latour, Law
Key problems of scientific approaches	Division of labour and the welfare state (population metrics)	Propaganda and media influence (audience metrics)	Science and technology (scientometrics)
Technical conditions	Hollerith's machine (tabu- lating calculation)	Radio, telephone	Internet, the Web, big data
Semiotic formats	Crosstabs, topographic maps	Curve and bar graphs, pie charts	Graphs, timelines, dashboards
Metrics	Statistics (classic)	Sampling	Machine learning
Technical criteria for data quality	Relevance, accuracy, time- liness, accessibility, com- parability, coherence	Confidence intervals, probabilities	Volume, variety, and velo- city (big data)
The social sciences' dominant modalities	Explanations	Descriptive and predictive correlations	Predictive correlations

38 *Dominique Boullier*

Brands may benefit from learning to react to using these metrics based on traces. Social sciences of “society” and “opinion” may also benefit in further developing their approaches by using these sources. In this sense, we plead to make these approaches coexist, to learn to change points of view, and to admit the conditions of possibility for each generation, relying on the states, media and brands. Each specific study of an issue arising from everyday experience or raised by prescribers, such as brands, must lead to a combination of the three generations. This is provided that research has a specific framework for these traces that invade our world. There is a new “raw” material that deserves a review of its own, and produces a third layer to the social measurable according to other principles, and not reducible to “society” or to “opinion”. “Society” ended up existing, “opinion” ended up existing, and “replications” must eventually end up existing in the same way. The time may have come for the “buzz” to be translated in scientific terms and to gain some recognition of its own agency.

Notes

- 1 Cardon (2013) has proposed a typology consisting of links, clicks, likes and traces.
- 2 The works of Loïc Blondiaux (1998) and Jean Converse (1987) develop this story extensively.
- 3 Even though Kaier tested them in 1891 and Bowley established the principles of the probable error in 1912.
- 4 The Chinese equivalent of Twitter.
- 5 The original term used was “répliques”, which could also be translated as “aftershocks” or “tremors”.

References

- Blackmore S. (1999). *The meme machine*. Oxford: Oxford University Press.
- Blondiaux, L. (1998). *La fabrique de l'opinion. Une histoire sociale des sondages*. Paris: Le Seuil.
- Boullier, D. (2013). “Plates-formes de réseaux sociaux et répertoires d'action collective”, in Najar, S. (ed.), *Les réseaux sociaux sur internet à l'heure des transitions démocratiques*. Paris: Editions Karthala, 492 pp.
- Boullier, D. (2016). *Sociologie du numérique*. Paris: Armand Colin.
- Boullier, D. and Lohard, A. (2012). *Opinion mining et sentiment analysis. Méthodes et outils*. Paris: Open Editions Press.
- Bowker, G. (2014). “The Theory/Data Thing. Commentary”, *International Journal of Communication*, 8, 1795–1799.
- Boyd, D. and Crawford, K. (2011). “Six Provocations for Big Data”, Paper Presented at the Oxford Internet Institute's A Decade in Internet Time: Symposium On the Dynamics of the Internet and Society. Oxford: University of Oxford.
- Bruno, I. and Didier, E. (2013). *Benchmarking. L'État sous pression statistique*. Paris: La Découverte, Coll. Zones.
- Callon, M. and Latour, B. (1981). “Unscrewing the Big Leviathan: How Actors Macrostructure Reality and How Sociologists Help them to Do So”, in Knorr,

Big data in social sciences and market research 39

- Karin and Cicourel, A.V. (eds), *Advances in Social Theory and Methodology: Toward an Integration of Micro and Macro Sociologies*. London: Routledge and Kegan Paul.
- Cardon, D. (2013). “Du lien au like sur internet. Deux mesures de la reputation”, *Communications*, 93, La réputation, 173–186.
- Cochoy, F. (1999). *Une histoire du marketing. Discipliner l'économie de marché*. Paris: La Découverte.
- Callon, M., Law, J. and Rip, A. (1986). “Qualitative Scientometrics”, in Callon, M., Law, J., Rip, A. (ed.), *Mapping the Dynamics of Science and Technology*. London: Macmillan, pp.103–123.
- Converse, J. (1987). *Survey Research in the United States. Roots and Emergence 1890–1960*. University of California Press.
- Dawkins, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press.
- Desrosières, A. (1998). *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge, MA: Harvard University Press.
- Desrosières, A. (2014). *Prouver et gouverner: une analyse politique des statistiques publiques*, ed. Emmanuel Didier. Paris: La Découverte.
- Driscoll, K. and Walker, S. (2014). “Working within a Black Box: Transparency in the Collection and Production of Big Twitter Data”, *International Journal of Communication*, 8, 1745–1764.
- Durkheim, E. (1897). *Le suicide*. Paris: Alcan.
- Eisenstein, E. (1983). *The Printing Revolution in Early Modern Europe*. Cambridge: Cambridge University Press.
- Hannerz, U. (1983). *Exploring the City*. NY: Columbia University Press.
- James, W. (1890). *The Principles of Psychology*, two vols. Dover Publications.
- Jenkins, H., Ford, S. and Green, J. (2013). *Spreadable Media. Creating Value and Meaning in a Networked Culture*. New York and London: New York University Press.
- Kleinberg, J. (2002). “Bursty and Hierarchical Structure in Streams”, Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining.
- Kleinberg, J., Gibson, D. and Raghavan, P. (1998). “Inferring Web Communities from Link Topology”, in Proc. of the 9th ACM Conference on Hypertext and Hypermedia (HYPER-98), pp. 225–234, New York, June 20–24.
- Latour, B. (2005). *Reassembling the Social – An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.
- Latour, B. (2011). “Gabriel Tarde. La société comme possession. La preuve par l’orchestre”, in Debaise, D., *Philosophie des possessions*. Paris: Les Presses du Réel.
- Latour, B., Jensen, B., Venturini, T., Grauwin, S. and Boullier, D. (2012). “‘The Whole is Always Smaller than its Parts’: A Digital Test of Gabriel Tarde’s Monads”, *British Journal of Sociology*, 63(4), 590–615.
- Leskovec, J., Backstrom, L. and Kleinberg, J. (2009). “Meme-Tracking and the Dynamics of the News Cycle”, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).
- Manovich, L. (2012). “How to Compare One Million Images?” in Berry, D. (ed.), *Understanding Digital Humanities*. New York: Palgrave Macmillan.
- Marres, N. and Weltevrede, E. (2013). “Scraping the Social? Issues in Live Social Research”, *Journal of Cultural Economy*, 6(3), 313–335.
- Michel, J.B. et al. (2011). “Quantitative Analysis of Culture Using Millions of Digitized Books”, *Science*, 331(6014), 176–182.

40 *Dominique Boullier*

Rogers, E.M. (1983 [1963]). *Diffusion of Innovations*. New York: Free Press.

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

Schütz, A. (1962). *The Phenomenology of the Social World*. Evanston, IL: Northwestern University Press.

Tarde, G. (1989 [1901]). *L'opinion et la foule*. Paris: Puf.

Tarde, G. (2001 [1895]). *Les lois de l'imitation*. Paris: Les Empêcheurs de Penser en Rond.