*PA*rtecipazione e *CO*nflitto
** The Open Journal of Sociopolitical Studies*
http://siba-ese.unisalento.it/index.php/paco
ISSN: 1972-7623 (print version)
ISSN: 2035-6609 (electronic version)
***PACO*, Issue 11(2) 2018: 448-471**
DOI: 10.1285/i20356609v11i2p448

Published in July 15, 2018

RESEARCH ARTICLE

THE POLITICS OF BIG DATA ASSEMBLAGES

Biagio Aragona
University of Naples Federico II

Cristiano Felaco
University of Naples Federico II

Marina Marino
University of Naples Federico II

ABSTRACT:
One way to study the politics of big data is the inspection of their assemblages. By opening up the "black boxes" of data assemblages, it is possible to reconstruct the choices, compromises, conflicts and agreements that contributed to the construction of a given datum. Leaning on in-depth interviews and focus groups with experts and specialists who work within three European data centers, we unveil the interdependence between social and technical aspects and between a series of internal and external actors, which all contribute to the data assemblage. Results show that communities of experts, technologies, stakeholders and end-users are entwined components that interact amongst them in a contingent and complex web of negotiations and constraints and frame what is possible, desirable and expected by data.

KEYWORDS:
Data Assemblage, Big Data, Symbolic power, Qualitative interviews, Data centers.

CORRESPONDING AUTHORS:
Biagio Aragona: aragona@unina.it

1. Introduction

Big data are characterized by power relationships between many human and non-human actors. One way to study the politics of big data is by inspecting their assemblages.

It is with the work of Bourdieu (1979) that symbols in the form of data were revealed in their political use. Bourdieu (*ivi*, 77) criticized the process of encoding as a part of the symbolic power expressed by the State, because "symbolic power is a power of constructing reality". He considered numbers that represent phenomena as the center of public debates and political action. In a different fashion, Foucault (1980) with his concept "power-knowledge" clarified that power is based on knowledge and makes use of knowledge and that, on the other hand, power reproduces knowledge by shaping it in accordance with its intentions. In addition, the systemic perspective of power by Niklas Luhmann (1984; 1997) considered information and communication as strategic for power, intended as a process that improves the system organization and integration.

Alonso and Starr wrote in the *The Politics of Numbers* (1987, 4) that whenever the link between data and power is studied, the focus of analysis is on the "system for the production, distribution and use of numerical information". The importance of focusing on the systems that produce and frame data has also been advocated by a recent plurality of research inspired by Foucault's works, and it has merged into the field of critical data studies (Dalton and Thatcher 2014; Iliadis and Russo 2016). These studies aim to interrogate all forms of potentially depoliticized data science to track the ways in which data are generated, curated, and how they permeate and exert power. According to Kitchin and Lauriault (2014, 6), the subject of critical data studies should be the sociotechnical "data assemblage" that makes up big data, defined as a "complex sociotechnical system composed of many apparatuses and elements that are thoroughly entwined, whose central concern is the production of data". Furthermore, as noted by Kitchin, data assemblage "includes all of the technological, political, social and economic apparatuses that frame data" (2014, 26). The apparatuses interact with and shape each other through a contingent and complex web of multifaceted relations. As data are a product of the assemblage, the assemblage is structured and managed to produce data (Ribes and Jackson 2013). Data and their assemblage are thus mutually constituted, bound together in a set of contingent, relational and contextual, discursive and material practices and relations. Moreover, each data assemblage forms part of a wider datascape (Aragona and De Rosa 2018) composed of many other inter-related and interacting data assemblages and systems. The diffusion of the term assemblage,

in French *agencement*, is attributed to the French philosopher Deleuze. He believed that assemblages have the function of dismissing the representative thought that arrogates the control of metadiscursive knowledge, of disciplinary specialisms and related institutions. Assemblage is above all the attitude to recognize the production of symbols as fields of force in the entity in which they are located, and which they contribute to produce (Deleuze and Guattari 1980). The choice of this angle makes data assemblage an operational concept that focuses on the processes of imbalances and rebalancing between legal, economic, technological and social dynamics (Sassen 2008).

Given that data are the combined product of different apparatuses, layered analytic techniques, and various competing communities of experts, their origins and interpretations become multiple and conflicting, with the result of their assemblage being "black boxed". That of "black box" is a term used by cybernetics when a part of a mechanism or a series of instructions are unknown, apart from their inputs and outputs. As Pasquale (2015) notes, the term "black box" is a useful metaphor, given its dual meaning. It can refer to a recording device, like the data-monitoring systems in planes, trains, and cars. Or it can mean a system whose working logics are opaque; we can observe its inputs and outputs, but we cannot tell how one becomes the other. According to him, we face these two meanings daily: tracked ever more closely by firms and government, we have no clear idea of just how far much of these data can travel, how they are used, nor we can know their consequences. Nevertheless, "black boxing" data has always been a problem, even before the advent of big data. Data, no matter how big, are mobile immutables (Latour 1987), something that has its own stability, but it is part of a series of elaboration and exchange processes. Whenever black boxes are opened, the elaboration processes are revealed and problems, working groups, decisions, competitions and controversies disclose (*ivi*). By unpacking data assemblages, it is therefore possible to reconstruct choices, compromises, conflicts and agreements - the politics - which contributed to the production of a given datum.

However, it is somehow problematic to empirically identify data assemblages, because they have fluid and undefined boundaries, and each assemblage is inextricably linked with other data assemblages. In order to study them, we have therefore selected three European centers of calculation, which produce, use and share digital data: the Web Science Institute (WSI), the Italian National Institute of Statistics (ISTAT) and the Norwegian Center for Research Data (NSD). These centers are venues where all the apparatuses which form the data assemblage and can be viewed as a "vantage point" from which to better understand the politics of big data.

First of all, the activities of the three centers are influenced by governmentalities, political economy, finance and the market. For example, ISTAT introduces in the annual

program the statistical activities that are demanded by national and local governments for designing policies, and NSD is founded by the Research Council of Norway with the aim to facilitate the access to data for research. In the same fashion, WSI is sustained by partners in government, business and industry, establishing active engagement with the Web community and governance bodies. Moreover, they are organizations and institutions where a certain number of practices takes place based on some forms of knowledge and system of thoughts that are shared within the communities of experts who participate in the activities of the center. For instance, ISTAT is a research institute that employs routinized activities (i.e., the Generic Statistics Business Process Model) that are built upon some conventions existing between the communities of statisticians, IT and domain experts. NSD adopts the Statistical Data and Metadata eXchange (SDMX) protocol that aims at standardizing the processes for the exchange of statistical data and metadata among international organizations and their member countries. WSI is an honorary founding partner of the Open Data Institute, a world-leading organization that pioneers new social and commercial value from open data. Finally, they are based in specific places and their activities are interrelated to the infrastructures that are connected to the centers. The budget of ISTAT, mainly coming from European Commission and the Ministry of Internal Affairs, impacts on the choice made in the institute about the kind of data collection activities that are set up, and the experimentations that are made. The activities realized are also strictly connected to the network of local statistical offices that run the data collection activities in practice; they form the National Statistical System (SISTAN) infrastructure, which is guided by ISTAT. NSD is a Limited Company owned by the Norwegian Ministry of Education and Research and at the same time is service provider to CESSDA (Consortium of European Social Science Data Archives), which is acknowledged by EU as the only European data infrastructure in the field of social sciences. Similarly, the WS draws together world-leading researchers from across the University of Southampton and it is part of the Web Science Network of Laboratories (WSTnet), an International network bringing together world-class research laboratories to support the Web Science research.

We questioned experts and professionals who work within these three European data centers by means of focus groups and in-depth interviews, to unveil useful information about the political, technical and cultural aspects of big data assemblages. More specifically, we conducted in-depth interviews with directors and head of sections of the centers to elicit their critical reflection on data assemblage and its apparatuses. In addition, we ran focus groups in each center involving team members without managerial responsibilities. Thus, focus groups allowed for a simulation of the whole of

activities, choices, relations and negotiations related to the data assemblage taking place within centers.

The paper is structured as follows: section 2 aims to define the linkage between data and politics; section 3 frames the research design; section 4 presents the main results of the analysis of the interviews conducted within the three European centers of data calculation. The last section concludes with some future perspectives from which to continue the critical analysis of data assemblages.

2. The politics of big data

The linkage between data and politics is not new. It is at least since the foundation of the European nation states that knowledge in the form of data represented an instrument of power. Statistics, from their very beginning, combined "the norms of the scientific world with those of the modern, rational state" (Desrosiéres 1998, 8). The traditional approach to data was based on the assumption of independence, between measured reality and a measurement process: data were considered to be able to 'objectively' embrace the analyzed phenomenon. Along with objectivity there was pragmatism; data were not needed for knowledge alone, but for administrative and political goals. Every piece of data was valuable and reliable only if it was useful in practice: a means to an end. In this context, data were the factual terrain upon which judgement (Sen 1990) was directly dependent, and this way of looking at data has lasted until today, for example through the evidence-based policies model (Stoker and Evans 2016).

Data do not exist prior to social action, but through social action (Bowker 2013) nor do they exist independently of relational processes. Rather, they are the product of choices and constraints constituting systems of thought, technologies, people, resources and funding, know-how, public and political opinion, ethical considerations that together affect the processes of production, management, sharing and analysis of data (Bowker and Star 1999; Lauriault 2012; Ribes and Jackson 2013). Similarly, data assemblage is not simply a neutral system. Data and their assemblage are situated, contingent and relational, thus co-determined and mutually constituted, and employed in order to achieve certain aims (Poovey 1998; Latour 1987; Hacking 1982; Anderson 1991).

A rich field of study, starting from the 1980s, and influenced by the works of Bourdieu (1988, 1991) on State and symbolic power, and of Sen (1990) on informational basis of judgement, capabilities and choices, began to question the traditional objectivist and pragmatist vision of data. In such a new context, data were seen as the results of

deliberate processes of choice, selection and justification – in other words, as the outcome of a political process (de Leonardis 2009). Much research (Thévenot 1984; Alonso and Starr 1987; Desrosières and Thévenot 1988; Salais and Storper 1993; Desrosières 2010) has then been devoted to studying the processes where classifications, indicators and measures, and the data they generate are constructed through a series of conflicts, compromises and agreements between many actors with different cognitive frames. Data then revealed their conventional nature, or to say it with Thévenot's (1984) words, data became "agreed" within a specific format that is given in a particular action and justification regime. Hacking (2007) showed that scientific knowledge and expert skills participate in these political processes where objects and data are created, and they become also the main field where this political process concretizes. In this perspective, censuses, indexes, indicators, registers, catalogues, archives are just an outcome of the socio-political process of making, of the "politics of indicators" (Salais 2004).

Data do not happen through unstructured social practices "but through structured and structuring fields, in and through which various agents and their interests generate forms of expertise, interpretation, concepts, and methods that collectively function as fields of power and knowledge" (Ruppert *et al.* 2017, 3). This means that the existence and definition of data should be seen as conventions subjected to debate. The final question is therefore if these processes are black boxed or if, on the contrary, choices, compromises and agreements are visible and publicly questionable, and if the construction of data can be overhauled. Many scholars have researched how citizens have challenged the social categories of data regimes and their effects (Anderson and Fienberg 2000; Kertzer and Arel 2002; Nobles 2000), but these analyses refer to a specific data regime which was public-centered. Indeed, as noted elsewhere: "the state, or rather organizations, institutions, agencies, agents, and authorities that make up the complex field of government, maintained an effective monopoly on data regimes" (Ruppert et al 2017, 3).

This scene, where institutions and agencies had the monopoly on data production and collection, has been increasingly challenged by the advent of big data. Currently, a great variety of actors is starting to play key roles in the data governance. To be sure, since the introduction of the principle of subsidiarity in the production of statistics, the networks of agencies devoted to the construction of data has widened, with the entrance of various international organizations such as the United Nations, the European Union, OECD, and ILO. But these networks were perpetuating the same data regime that was created when the State had the monopoly on data production. On the contrary, with the increasing gathering and deployment of data by corporations (Thrift 2005),

the data assemblages have changed sharply. The main reason is that the value associated to data has changed. In a data regime governed by public actors, data were considered as a public good. As the value of public goods is inverse to their scarcity, the more the good is diffused, the higher its value. Conversely, in the private market scarcity gives value to the good. Thus, the rarer the good, the higher its value. This means that privatization of data has completely subverted the linkage between data and value, and we can also add between data and power. Since then, big data have been produced by the major corporations in communication and logistics, but, at same time, they have been employed for governmental and administrative purposes. A new balance of power between public and private actors has been replaced.

Thus, big data have clearly become a political terrain characterized by strong power dynamics between private and public actors. Each big data assemblage represents a political terrain where a series of networks between different actors is established. These networks are mixed public and private, multi-level (related not only to the single institution which holds the data, but also to the others connected to it at the national and international level) and multi-stakeholder. Different end-users are involved, including individual actors (citizens, civil servants, beneficiaries, etc.) and collective ones (local governments, national government, private companies, NGO, etc.).

Some features of the politics of these new big data assemblages can be summarized as follows. A first aspect is the growth of data intermediaries, which are claiming the authority on production and dissemination of quantitative information. They mash up data with multimedia contents and comments and are able to reach a wide audience through digital devices. Data brokers are an example in this respect, since they aggregate and analyze online data, and then interpret the results of these analyses. They make profit by selling the reports of these analyses to interested companies or agencies. A further example of how big data are reconfiguring the relations between private and public data actors are *data collaboratives*, an emerging and increasingly common form of public-private partnership in which actors from different sectors exchange data to create new value. Such collaborative arrangements, for example between social media companies and humanitarian organizations or civil society actors, are often built on more than simply the exchange of data as the cross-sector exchange of expertise, knowledge, and resources also play a key role in achieving success (Verhulst and Young 2017).

The question of expertise and skills is also crucial for understanding the politics of big data assemblages. Kitchin (2014) noted that the management of vast amounts of continuous data is a technical challenge that public bodies are not well equipped to face. Similarly, Emanuele Baldacci, current director of methodology at Eurostat, in 2016

wrote that the majority of people working for statistical offices do not have the proper skills needed for handling big data (Baldacci 2016). This seems to indicate that the assemblage of big data has had a strong impact on the communities of experts who have been dealing with data and their power for a long time: statisticians, experts, and computer scientists have all had to renew their skills according to big data.

A further political point of big data assemblages is the power struggle between all the different stakeholders that in some way are connected to and interested in data and in their use. While traditional data assemblage stopped when data were released, current big data assemblages must follow up on the way data are handled by final users through the media and then try to understand if users are able to transform these data into knowledge, and how this process works (Giovannini 2014). Data are not useful in and of themselves. It is what is done with data that is important. As Conte (2016) stated, big data may be an opportunity for social sciences and societies under the condition that the application of computation to social phenomena will be oriented to policy making. In the same fashion, Supiot (2016) reminds us that *La gouvernance par les nombres* may be crucial to understand what the future of big data will be within both social sciences and society. But because the users of big data are many and various (administrators and policy makers, politicians, business companies, researchers, journalists, citizens), and any of them has its own informational needs that may be partly in conflict, the balance between the needs of the different stakeholders is matter of power and struggle - another form of the politics of big data.

3. Researching big data assemblages

Unpacking data assemblage means delving into various aspects of three main domains: things (infrastructures, devices, techniques, etc.), language (code, algorithms, etc.) and people (scientists, users, etc.). The complex nature of its apparatuses makes it difficult to isolate and then analyze its various aspects. Therefore, researchers are faced with the need to pinpoint the space where these elements may be observed. That is exactly why we chose to conduct our research looking at the centers of data calculation. These centers, indeed, can be viewed as venues in which all the apparatuses that constitute data assemblage converge. As mentioned above, the three centers are the Web Science Institute (WSI), the Italian National Institute of Statistics (ISTAT) and the Norwegian Center for Research Data (NSD). We chose these centers specifically because they act in different contexts, with very different missions and organizational structures (Tab.1).

Table 1 – Characteristics of the centers

Name	Country	Mission	Organizational structure
Istituto Nazionale di Statistica (ISTAT)	Italy	To serve the community by producing and communicating official high-quality statistical information.	Highly hierarchical. Departments, Sections, Units
Norwegian Center for Research Data (NSD)	Norway	To handle data services to institutions in order to improve the control and quality assurance of their own research data.	Hierarchical Departments only: IT, Data services, Data protection.
Web Science Institute (WSI)	UK	To undertake interdisciplinary research that can lead government, business and civic engagement to maximize the impact of Web technologies.	Flat, Directors and staff, no levels of middle management.

Source: Authors' elaboration

ISTAT's mission is to serve society by producing and communicating official high-quality statistical information in Italy, and it has a highly hierarchical organizational structure. NSD has a less hierarchical structure than ISTAT, composed only of sections and not of departments, and it has the mission to handle data services to institutions in order to improve the control and quality assurance of their own research data. Finally, WSI, which has a flat organizational structure without levels of middle management, aims to undertake interdisciplinary research that can lead government, business and civic engagement to maximize the impact of Web technologies.

Beside having different missions and organizations, the three centers act in distinctive contexts that may influence the various apparatuses of the assemblage. However, they share some common political features that make them comparable in this respect. First of all, they employ experts who manage skills and background knowledges coming mainly from computer science, statistics, social science and law. Furthermore, they relate with various stakeholders and end-users who influence and, at the same time, are influenced by their activities (i.e. national and local governments, corporations, politicians, administrators, citizens, researchers, journalists, etc.). Finally, they regularly face technical as well as ethical problems in the production, management and analysis of data.

The analysis of data assemblages is commonly realized through ethnographies (Geiger 2017, Seaver, 2017). We decided instead to interview experts and professionals who work in these centers of data calculation and are directly involved in data assem-

blage. We aimed to understand the meanings and the importance that actors participating in the assemblage give to the activities they run, according to their roles, background knowledge and the centers they work for. Interviews encouraged subjects also to critically reflect on the various facets of the assemblage, allowing us to understand their level of engagement in the different processes carried out by the centers.

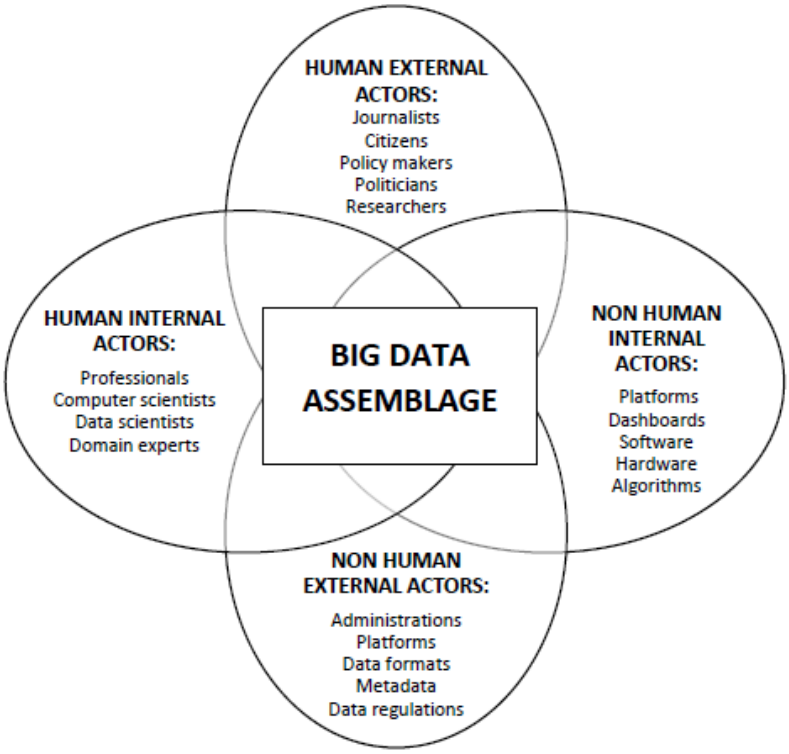
Field activities were carried out in the period between October 2017 and February 2018. We conducted in-depth interviews with two directors and six head of sections of the centers. More specifically, we interviewed the directors of the centers if the centers had a less complex structure, and the head of sections if the centers were bigger and more hierarchical. Different from structured interviews, in-depth interviews encourage subjects to talk with greater autonomy, fostering a critical reflection on apparatuses and eliciting a reconstruction of data assemblage during the interview. To this aim, we prepared an interview guide in advance that was composed of a list of topics about of the general work of the centers to be covered over the course of the interview.

Along with interviews, we conducted three focus groups, one for each center, with data team members without managerial functions. More specifically, we selected experts working in the following sections: information technology, methodology, legal and data protection. The participants had therefore different educational and professional backgrounds and included computer scientists, social and political scientists, statisticians and legal experts on data protection. Compared to interviews, focus groups allowed us to get a wider range of views and to explore various procedures from different perspectives. In this case, we created a more structured schedule than that we employed during interviews and which included a short list of predetermined questions on the meaning assigned to the specific practices of the center. By simulating the relational dynamics that take place within data assemblage, focus groups informed us on agreements and disagreements between the different communities of experts and on their level of engagement in the different stages of the assemblage.

Data assemblage is composed of many human and non-human actors that are internal and external to data centers. The power relationships between these actors materialize within the different entwined apparatuses. In this paper, we have selected only the parts of the interviews about the political aspects that interviewees highlighted in describing the elements that constitute data assemblage. We then classified the selected materials around four main topics. Firstly, we analyze the skills requested for working with big data and the relationship between various domains of expertise. Furthermore, we disentangle how technical problems can hinder the whole process of data construction, management and analysis, and what strategies are adopted to face

them. Another main topic concerns the need to merge the interests of the many end-users and stakeholders (fig.1), stepping into negotiations and compromises in order to ensure data quality. Finally, we address ethical implications within data assemblage and more specifically the difficulty to conciliate all the different ethical regulations of the various data sources employed.

Figure 1 – The actors of Big Data Assemblage



Source: Authors' elaboration

4. Results

4.1 Interdisciplinarity

A first point that interviewees mention is about the communities of experts that participate in the processes of production, management and analysis of data. The areas of expertise needed are related to domain, data and computation, and it is essential to combine skills and viewpoints that cut across disciplines. A dialogue between the different communities is required to blend methodologies and disciplinary matrixes, and shape what Lackatos (1976) called "background knowledge" – i.e., the whole set of facts and parameters used in the construction of any given theory, and of any given data.

First of all, interviewees refer to various styles of doing research with big data. From their points of view, interdisciplinarity is seen as a valid research model to be pursued that enriches the quality of the data produced and the understanding of the phenomena of interest. This model is realized through cooperative moments of design and training. This point is well expressed by the words of one of our interviewees:

> *All the interdisciplinary work is a lot of co-design […] with [Name of project] a lot of work is just sitting in a room talking through problems. […] A lot of our discussions have been about prioritizing things and saying "well we can live without that, so we won't get you to do that", "this is really crucial and I know it will take a bit of time, so do that" or "this isn't massively important but you are telling us it will only take a day, so actually let's have that as well". So, there's a lot of that sort of negotiation. (M., WSI)*

However, interviewees substantially underline the presence of strong boundaries between the different communities of experts, which constrains the implementation of interdisciplinarity. In addition, they highlight the need for specific skills to work with big data, which are not simply technical, but also deeply epistemological, and take form in the ability of mixing social theory and computation, data and modelling in an innovative way. In this respect, some interviewees highlight the difficulty that they have in finding experts with these skills on the labor market; they claim that the education system needs to be more focused on targeting big data, with the aim of creating future professionals specialized in this field. The lack of these kind of professionals forces the centers to adapt their own internal resources or to turn to external actors for assistance and training. As one of the researchers from ISTAT pointed out:

We have bought Cloudera platform, and at the moment the serious problem is that IT staff is still under training; we have also outsourced a service training on big data for IT staff, but we would like to include also the staff of the production and method sections, in order to develop IT competence […] and to enable it to work in complete autonomy. […] We're really starting to re-examine the training and acquisition of professionals, also from abroad […] I think that university should provide more competences to the students to work with these kinds of data (G., ISTAT)

4.2 Non-human actors

Interviewees underline that non-human actors (such as software, platforms, etc.) can become constraints that intervene in the whole process of data construction, management and analysis. First of all, some interviewees admit the difficulty in keeping up with the technological changes when they perform web scraping activities, because the website structures are constantly undergoing changes (Lieberman, 2008). The velocity and ever-changing nature of big data generates acquisition problems, and it requires specific technical skills and technological equipment to perform customized data capturing strategies able to follow changes of platforms and of the Application Programming Interfaces (API). Some actors can get through these problems with the development of specific tools, like in the case of a WSI a researcher who needed to write their own software to obtain a dynamic visualization which showed networks growing over time:

I wanted a dynamic visualization which showed the network growth over time. […] I spent a few weeks trying to make Gephi software work […] and I actually set aside three weeks and managed to put three weeks in my diary where I was just gonna focus on this. […] And so, what I realized was that I was going to need to write my own software to do exactly the visualization that I wanted. (L., WSI)

Furthermore, one of the claims about the data revolution is that it is possible to create datasets with strong relationality, which can then be combined to generate additional insight and value (Mayer-Schyonberger and Cukier2012). For data to be integrated into new datasets, several elements are required – such as shared indexical fields and data standards, consistent metadata systems and compatible format, such as SDMX used for disclosure authority in Europe, and DDI adopted in the social science field. But that it is not always case. For example, when describing a scientometric pro-

ject which involved the merge of many databases coming from different institutions, one interviewee expressed the difficulty to conciliate all these dissimilar standards.

> *The data sources that I use have been 2000-3000 institutional repositories around the globe […] with their different uses of the different standards. There are, say, 3 or 4 major platforms... but each of those...have 10 different versions around... and they use different metadata standards… And then you've got the different archival and librarian practices in every institution and they'll use the software differently, and use the metadata alternatively. (P., WSI)*

Another interviewee further specifies the problem as follows:

> *We've been using various metadata… DDI, for instance…. SDMX… that's a system that uses a lot of space for European disclosure authority. It's very oriented towards communication and transport: transport is a tabulated data for instance, a very statistical authority type; DDI, data documentation initiative, is much more into social sciences. We have used them quite extensively because we developed a kit of software. (A., NSD)*

Finally, not only can the software interfere with the final assemblage, but also the decisions that are made to construct dashboards or platforms may have great impact on the kind of data that are generated, on their quality, and possible uses for further analysis and insights. As an example, interviewees explain that, when designing a dashboard, they have constructed an incorrect variable that undercounted the access to the dashboard. In other cases, technical restrictions have to be dealt with stakeholders' demands, as in the case of the construction of a platform where a number of different export functions was developed just to meet the demands of the psychologists:

> *Over time we've developed a number of different export functions and we've gone through a number of sort of co-design processes with the psychologists to find out what forms of data they want and how […] when we were talking about the data and exporting the data, it either had to be simple, so .csv files, that are easily changed to do whatever they want, or SPSS, because they have a tradition of using SPSS. (M., WSI)*

4.3 Links with external actors

A first issue about the link between internal and external actors is connected to access to the data sources. Many different sources may be employed in the assemblage coming from institutions, corporations and data brokers, with different standards about metadata, operational definition and data structures. Accessibility is a fundamental issue, and in recent years, at least for data generated by public funded research or by public agencies, some access criticalities have been overcome through open data initiatives and the building of data archives and data infrastructures aimed to sharing and making data available for analysis.

Access becomes harder with respect to data produced by private companies because they are under no obligation to share the data they produce for handling their services. Taylor et al. (2014) argue that access to corporate data may address research results, because of their proprietary nature which may limit the replicability of studies. Access to data is usually individually negotiated and it involves signing a series of agreements concerning intellectual property, non-disclosure and re-sharing. In some cases, a selection of data may be available through API. In some cases, difficulties to set necessary negotiations to get data may compromise their quality, especially with regard to social media data. As a researcher from ISTAT argues:

> *You can imagine the effort to get detailed records; agreements between institutions and authorities, and then with the guarantee authority […]. I spent two years trying to obtain contacts, appointments and agreements. […] (A., ISTAT)*

Some of the interviewees underlines an overall difficulty to access also social networks data sources, stating that not all social network platforms offer the same data quality. They think that Facebook is the richest data source but it's no longer accessible, while the collection of data from Twitter is free, but often "massively irritating because of the constraints". Accessing to these kinds of data may require a "special relationship" with these companies. In this perspective, the figure of data brokers is particularly important as they allow for the acquisition of a large amount of data and layers of services. More specifically, data brokers are often companies (data aggregators, consolidators and resellers) that capture, gather together and repackage data into privately held centers of data calculation for rent or re-sale on a for-profit basis:

> *We've got a range of channels for getting social data. One of those to get through is by paying an intermediate company that gathers social data and provides some add-*

ed value analysis. […] the existence of data brokers that had their important role and their partners because of the connections that would make commercial models allows us to buy a lot of the data and layers of services. (S., WSI)

Another issue pertains to end-users, who can be of many types – e.g., policy makers, researchers, citizens, communication experts as data journalist – and with different goals. Interviewees are usually in contact with public institutions, such as universities and industrial partners to whom they do not only provide services but also often construct long-lasting partnership by sharing common projects. In this respect, the words of two researchers, one from the WSI and one from the NSD are quite telling:

There are a lot of academic partners, and then there are the industry partners and government. […] we are connected with all the big public services, so the education service, the fire service, the police, City Council […]and we are going to do some collaborative projects with them with the idea not only of producing something useful for them, but also by collaborating with them. (L., WSI)

We're still working closely with the national research council […] and then we gather all the researchers, maybe with universities and educational systems we run a lot of informational activities, telling students and researchers that we exist and the possibility we give. (B., NSD)

Moreover, it is quite interesting how the interviewees are aware of the negotiations needed to merge the interests of the different end-users and stakeholders, and to reach a compromise between the parties. Sometimes centres carry on collaborative projects that not only produce useful results for stakeholders, but they also are a way to set up some forms of collaborations between them. Furthermore, centres organize events where different end-users may receive information about their activities and about the services they can offer to different targets of users.

I think we are quite used to the idea of putting different views on data for different users' groups. Hopefully, would be as a negotiation. You can imagine if you go on different stakeholders, you might want them all investing in the code design, on what you are building, and so there were having to be compromises […] they acknowledge that and they are aware of that and they see that it is trying to solve multiple problems at the same time. (M., WSI)

4.4 Ethical implications

A main final aspect where the politics of big data clearly emerge is the ethical implications of data assemblage. Data are generated and employed for many ends including governing societies, managing organizations, leveraging profits, and regulating places. The generation of data and the work done with the are inherently infused with ethical concerns that in turn refer to data preservation, data security, consent and privacy.

First of all, interviewees talk about the ethic of disciplines: ethic cannot be seen in absolute terms, it is in fact described as a relative concept that changes among various disciplines. This aspect is particularly evident in the following excerpt from one of our interviews at WSI:

> *It is really difficult to understand the differences between the ethics in the different disciplines, so in computer sciences until very recently at least, if something is in the public domain you don't need an ethic approval for it, in social sciences you do. (S., WSI)*

Furthermore, independently of how ethics are conceived, interviewees believe that ethical concerns are more urgent with social media data, because terms and conditions about the processing and use of these data are sometimes less restrictive. On the contrary, the use of data in public sector is often more limited than the private one. For example, in Italy the use of call detail records is permitted only once the appropriate safeguards to the guarantee authority have been given. In addition, the internet of things and the wide diffusion of sensors does not always provide enough information to allow a full understanding of the processes and the goals of personal data collection and analysis. This practice challenges the entire ethical system that has been created and institutionalized upon survey data, a kind of data that is different from big data.

A final emerging need is the difficulty to conciliate all the different ethical regulations of the various data sources employed. An interviewee highlights the problem of jointly using many data sources, as they can be regulated by divergent ethical guidelines and standards practices. What emerges is an “infraethic” (Floridi 2013), a hypernetworked ethics where the agents in the network may cause collateral consequences on all the others.

> *It's not only data protection issue connected to the GDPR, but also terms of use and a bigger issue of data ownership. When you collect the data you can get the legal consent to collect the data, but what about the data from Facebook? [...] You can get it*

from companies that are set up with Facebook […] data protection is not very good, there is a protection, but you pay for it… (M., NSD)

5. Conclusions

The politics of big data surface at least at three distinct levels of study: the disciplines, the data sources and the administration of ethical issues.

The importance of establishing a dialogue between different communities of experts is considered by all interviewees as central for developing a data culture (Aragona 2008) within big data assemblage, but respondents believe that some obstacles should be still overcome. Big data indeed threaten to divide scientists into antagonistic methodological camps built for example around access. Access may be granted only to some according to their influence, budget and goals. As Boyd and Crawford point out, "this produces considerable unevenness in the system: those with money – or those inside the company – can produce a different type of research than those outside" (2012, 674). Furthermore, it clearly emerges that new information resources are altering the balance of power among different institutions and between the public/private and not-for-profit/for-profit sectors. The question of the difference between data-rich (institutions and corporations) and data-poor (scholars, researchers) actors of the assemblage is an interesting one. According to the interviewees, data rich actors, mainly the private ones, are more concerned about consolidating their competitive advantage than about improving data quality and data access. Finally, interviewees believe that big data require disruptive innovations in the way ethics is bureaucratized. In "old" data assemblages, generators, collectors and users often interacted, so there was room to negotiate consent, property, sharing and re-use. The various actors involved in big data assemblage, in order to adequately address ethical concerns, have to deal with the specific way technology works on every singular big datum. Social media data, sensors, transactional data and the techniques needed to extract, manage and analyze them (i.e. web scraping, data mining, machine learning, etc.), all pose different ethical problems to be overcome. There is often no incentive to develop mechanisms to inform the data subjects that their data will be used to support a different cause. In this debate, we may find on one side actors who demand stricter accountability, even if it means not exploiting big data to its full potential, and, on the other side, actors who believe that because it is on-line, using the data is inherently ethical.

Our analysis brings with itself three sets of implications: epistemological, methodological and normative. If big data are political, and their final form is defined by the

choices of many different actors, and by the constraints (economic, technical, organizational, etc.) which limit the activities of these actors –despite popular rhetoric to the contrary (Anderson 2008) –there is no such thing as "naturally occurring data" or "raw data" (Gitelman 2013). Big data are not simply a mass of empirical evidence that guides decision making processes, but socio-technical constructs that must be studied when they are in action. Unbundling the data assemblage is the only way to ensure transparency and quality and, thus, to follow the political use and the agency role of data.

Secondly, the analysis of interviews allowed us to unveil black boxed aspects of big data assemblage, fostering the understanding of the paradoxes of infrastructure as both transparent and opaque (Star 1999). We believe that more research is needed on the analysis of big data teams and on the reconstruction of big data "pipelines", and that this research must take advantage of the qualitative methods that are commonly used in social sciences.

A conclusive remark is normative and refers to the use of big data for the definition of public policies. The recognition of the political role of data, and of choices about data, is a significant step toward transparency, and the accountability of decision making. The intensive use of massive databases and the wide application of algorithms have raised political concerns, because they may lead to a technocratic form of governance (Mattern 2013). A further risk is that big data may accelerate a process of corporatization of the public arena. Big data are mainly private data coming from the largest software and hardware services companies and from the big majors of communication and logistic. While there is a wide range of evidence available, the challenge is to ensure that the selection of evidence used in policy-making is not only pertinent and relevant to current policy issues, but also transparent and ethical. Unpacking the data assemblages may be one way to increase the system's responsiveness and give politics a new source of legitimation, instead of reproducing new forms of technocratic regimes. At the same time, the awareness that data production may constitute fields of power, could lead policy-makers to support the decisions that they have already taken. They may do an even more accurate selection of only the data that justify their choices, without any form of negotiations. In this latter case, more research should be devoted also to understanding how data assemblages work within public administrations, and what is their level of readiness in taking advantage of the data they, and others, produce.

Finally, in addition to power dynamics, the analysis of big data assemblages may shed light also on counter-power. It is true that rationalization processes based on data and big data are one of the key elements that support the dynamic of global power,

but it is equally true that the scientific approach based on the same data has been a great social force of progress and emancipation. According to Antonelli, "the domain and its criticism, the ruling élites and protest movements have appealed to data" (2016, 360). Along with the expansive domain of instrumental rationality by governments and corporations, there is a possibility of an emancipatory rationality through data, as many cases of big data-activism (e.g., Milan 2017) have shown. Research within social movements, NGOs, and organizations that create new forms of collective action through data are another fragment of the politics of big data assemblage.

References

Alonso W., P. Starr (eds. 1987), *The politics of numbers*, New York: Russell Sage Foundation.

Anderson B. (1991), *Imagined Communities,* Revised Edition, New York: Verso.

Anderson C. (2008), "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", *Wired magazine*, 16(7): 16-07.

Anderson M., S. E. Fienberg (2000), "Partisan politics at work: Sampling and the 2000 census", *PS: Political Science & Politics*, 33(4): 795-799.

Antonelli F. (2016), "Ambivalence of official statistics: some theoretical-methodological notes", *International Review of Sociology*, 26(3), 354-366.

Aragona B. (2008), "Una nuova cultura del dato", *Sociologia e ricerca sociale*.

Aragona B., R. De Rosa (2018), "Policy making at the time of Big Data: datascape, datasphere, data culture", *AIS Journal of Sociology*: 173-185.

Baldacci E. (2016), *Innovation in statistical processes and products: a European view*, Dodicesima Conferenza nazionale di statistica "Più forza ai dati: un valore per il Paese", Rome, June 22-24, 2016.

Bourdieu P. (1979), "Symbolic power", *Critique of anthropology*, 4(13-14), 77-85.

Bourdieu P. (1988), "Social space and symbolic power", *Sociological Theory,* 7(1): 14-25.

Bourdieu P. (1991), *Language and symbolic power*, Harvard: Harvard University Press.

boyd d., K. Crawford (2012), "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon", *Information, communication & society*, 15(5): 662-679.

Bowker G., L. Star (1999), *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press, Cambridge.

Bowker G.C. (2013), "Data flakes: An afterword to 'Raw Data' is an oxymoron," in L.

Gitelman (ed.), *"Raw data" is an oxymoron*, Cambridge, Mass.: MIT Press, pp. 167-171.

Conte R. (2016), "Big Data: un'opportunità per le scienze sociali?", *Sociologia e Ricerca Sociale*, 109(3): 18-27.

Dalton C., J. Thatcher (2014), "What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data'", *Society and Space open site*.

de Leonardis O. (2009), "Conoscenza e democrazia nelle scelte di giustizia: un'introduzione", *La Rivista delle politiche sociali*, 2009(3): 73-84.

Deleuze G., Guattari F. (1980), *Mille Plateaux. Capitalisme et Schizophrénie*, Paris: Les Editions de Minuit.

Desroiséres A. (1998), *The politics of large numbers: a history of statistical reasoning*, Cambridge: Harward University Press.

Desroiséres A. (2010), *La politique des grands nombres*, Paris: Editions La Découverte.

Desrosières A., L. Thévenot (1988), *Les catégories socioprofessionnelles*, Paris: Editions La Découverte.

Floridi L. (2013), *The philosophy of information*, Oxford: OUP.

Foucault M. (1980), *Power/knowledge*, New York: Pantheon Book.

Geiger R. S. (2017), "Beyond opening up the black box: Investigating the role of algorithmic systems in Wikipedian organizational culture", *Big Data & Society*, 4(2), 2053951717730735.

Giovannini E. (2014), *Conoscenza e politica al tempo dei Big Data*, Bologna: Il Mulino.

Gitelman L. (2013), *'Raw Data' is an Oxymoron,* Cambridge: MIT Press.

Hacking I. (1982), "Biopower and the avalanche of numbers", *Humanities in Society,* 5(3-4): 279-295.

Hacking I. (2007), "Kinds of people: Moving targets", in *Proceedings-British Academy*, Oxford University Press INC., p. 285.

Iliadis A., F. Russo (2016), "Critical data studies: An introduction", *Big Data & Society*, 3(2), 2053951716674238.

Kertzer D., D. Arel (2002), *Census and identity*, New York.

Kitchin R. (2014), *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, London: Sage.

Kitchin R., T. P. Lauriault (2014), *Towards critical data studies: Charting and unpacking data assemblages and their work*, The programmable city working paper, 2.

Lackatos I. (1976), "Proof and refutations", in J. Worral and E. Zahar (eds.), *The logic of mathematical discovery*, Cambridge: Cambridge University Press.

Latour B. (1987), *Science in action*, Cambridge (MA): Harvard University Press.

Lauriault T.P. (2012), Data, Infrastructures and Geographical Imaginations. Ph.D. Thesis, Carleton University, Ottawa.

Lieberman D. Z. (2008), "Evaluation of the stability and validity of participant samples recruited over the internet", *CyberPsychology & Behavior*, 11(6), 743-745.

Luhmann N. (1984), *Soziale systeme*, Frankfurt am Main: Suhrkamp.

Luhmann N. (1997), "Limits of steering", *Theory, culture & society*, 14(1): 41-57.

Mattern S. (2013), Methodolatry and the art of measure: The new wave of urban data science, *Places Journal,* Retrieved November 5, 2013 (https://placesjournal.org/article/methodolatry-and-the-art-of-measure/).

Mayer-Schönberger V., K. Cukier (2012), *Big Data: A revolution that transforms how we work, live, and think*, Boston: Houghton Mifflin Harcourt.

Milan S. (2017), "Data Activism as the New Frontier of Media Activism", in G. Yang and V. Pickard (eds.), *Media Activism in the Digital Age*, London: Routledge.

Nobles M. (2000), *Shades of citizenship: Race and the census in modern politics*, California: Stanford University Press.

Pasquale F. (2015), *The black box society: The secret algorithms that control money and information*, Harvard: Harvard University Press.

Poovey M. (1998), *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*, Chicago: University Chicago Press.

Ribes D., S. J. Jackson (2013), "Data bite man: the work of sustaining long-term study", in L. Gitelman (ed.), "*Raw data" is an oxymoron*, Cambridge, Mass.: MIT Press, pp. 147-166.

Ruppert E., Isin E., and D. Bigo (2017), "Data politics", *Big Data & Society*: 4(2): 1-7.

Salais R. (2004), "La politique des indicateurs. Du taux de chômage au taux d'emploi dans la stratégie européenne pour l'emploi (SEE)", in B. Zimmermann, *Action publique et sciences sociales*, Paris: Editions de la Maison des Sciences de l'Homme.

Salais R., M. Storper (1993), "Les mondes de production (enquête sur l'identité économique de la France)", *Civilisations et sociétés.*

Sassen S. (2008), "Neither global nor national: novel assemblages of territory, authority and rights", *Ethics & global politics*, 1(1-2): 61-79.

Seaver N. (2017), "Algorithms as culture: Some tactics for the ethnography of algorithmic systems", *Big Data & Society*, 4(2), 2053951717738104.

Sen A. (1990), "Justice: Means versus Freedoms", *Philosophy & Public Affairs*, 19(2): 111-121.

Star S. L. (1999), "The Ethnography of Infrastructure", *American Behavioral Scientist*, 43(3): 377-391.

Stoker G., M. Evans (eds. 2016), *Evidence-based Policy Making in the Social Sciences: Methods that Matter,* Bristol: Policy Press.

Supiot A. (2016), *La Gouvernance par les nombres*, Paris: Fayard.

Taylor L, Schroeder R, and E. Meyer (2014), "Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?" *Big Data & Society*, 1(2): 7-16.

Thévenot L. (1984), "Rules and implements: investment in forms", *Information, International Social Science Council*, 23(1): 1-45.

Thrift N. (2005), *Knowing capitalism*, London: Sage.

Verhulst S.G., A. Young (2017), *Open data in developing economies,* New York: GOVLAB.

Funding

The research for this paper was funded by the research grant for original and innovative projects of the University Federico II of Naples.

Note by the Authors

The article is the result of the work of all authors, but Sections 1 and 2 can be attributed to Biagio Aragona, section 3 to Marina Marino, and the sections 4.1, 4.2, 4.3, 4.4 to Cristiano Felaco. The conclusions were jointly drafted by all authors.

Authors' Information

Biagio Aragona, Ph.D., is assistant professor in Sociology at the University of Naples Federico II and lecturer in Advanced Methods for Quantitative Research. He was member of the board of "Methodology" of Ais 2012-2015. His research activity is mainly on the use of statistical sources for the analysis of social policies and about the study of the challenges and opportunities that big data and others new data offer to social sciences.

Cristiano Felaco, PhD in Methodology of Social Sciences, is postdoctoral research fellow at the Department of Social Sciences, University of Naples with a project on Big Da-

ta Assemblages. His main interests of research are Text Analysis, Social Network Analysis, Big Data and Youth Transitions.

Marina Marino is Associate Professor of Statistics at the Department of Social Sciences of the University of Naples Federico II, where she is also a member of the research committee for the Ph.D. program on social science and statistics. Her chief research areas are computational statistics, data mining, classification and clustering, statistical analysis of interval-valued data and composite indicators.