Big Data from Below
Researching Data Assemblages

Biagio Aragona
University of Naples "Federico II"

Cristiano Felaco
University of Naples "Federico II"

Abstract: This paper aims to study the data assemblage of three centres of calculation which produce and use big data for social research. By unfolding how big data are produced we want to compare and contrast different aspects of data construction, management and exploitation. The results are drawn from focus groups and in-depth interviews of data experts. Respondents were interviewed about the activities of setting objectives, design decisions and choices with respect to expert languages, influences, constraints, debates with actors internal and external to data assemblage. The analysis presented in the paper focuses on the methodological activities run in the assemblage, and on subjectivities and communities involved in big data.

Keywords: data assemblage; big data; qualitative methods; data centres; data interoperability.

Submitted: July 18, 2018 – **Accepted:** May 20, 2019

Corresponding author: Cristiano Felaco, Department of Social Sciences, University of Naples "Federico II", Vico Monte di Pietà, 1 80138 Napoli (Italy). Email: cristiano.felaco@unina.it

1. Introduction

The article critically examines the data assemblages of three centres of calculation which produce and use big data for social research. The aim is to unfold how big data are produced by comparing and contrasting different aspects of data construction, management and exploitation. Furthermore, it addresses some criticalities in big data research in relation to contexts (public/private; national/international, etc.) and objectives (official statistics, policy design, academic research etc.).

Data are commonly considered neutral and objective material that condenses pieces of social reality in numbers and other symbolic forms,

but actually Manovich explains that: "data [do] not just exist, they have to be generated" (Manovich 2001, 224). In the philosophy of science, it is at least from the rise of post-positivist thinking that data have been critically considered a selection from the total sum of all possible data available (Kuhn 1962; Feyerabend 1969). Data are framed by methods and techniques, theories and background knowledges (Lakatos 1976), practices and contexts. Their production is situated and historically specific, a result of the conditions of inquiry, which are at once material, social and ethical. This idea that, to use the words of Gitelman (2013), "raw data is an oxymoron" (see also Leonelli 2016 for data in biology) raises questions about how data are assembled, and it calls for a critical investigation of the intertwined processes of collection, management and use that prepare data for becoming information, and then knowledge (Floridi 2010).

A long tradition of research has been devoted to study the processes where classifications, indicators and measures, and the data originate, are constructed through a series of activities where many actors with different cognitive frames interact (Thévenot 1984; Alonso and Starr 1987; Desrosières and Thévenot 1988; Salais and Storper 1993; Desrosières 2010). With the developments of data infrastructures, open data and big data, data intensive and positivistic approaches to scientific knowledge have disputed post-positivism (Kitchin 2014a). Discourses and practices surrounding the big data revolution (Mayer-Schonberger and Cukier 2012) moved towards an emerging variety of computational social science techniques (Lazer et al. 2009), which provide granular analyses that are said to no longer require theories (Anderson 2008) and critics (Iliadis and Russo 2016). The need to unpack big data assemblages has been then advocated by Dalton and Thatcher (2014), who have called for 'Critical data studies' (CDS), studies that apply critical social theory to data to explore the ways in which they are never neutral, objective, raw representation of social reality, but are situated, contingent, relational and contextual.

The objective of our research is to reconstruct contexts, activities and the long chain of human and non-human actors which construct big data. We interviewed experts and professionals who work within three European data centres by means of focus groups and in-depth interviews. We chose these interviewees because they are directly involved in big data assemblages and may reveal relevant information about its socio-technical apparatuses. The analysis focuses on three specific topics: some methodological challenges of data curation and data management that arise in a context of multi-stakeholder informational needs and objectives; the skills needed and the interdisciplinarity approach for dealing with big data; the ethical implications of using digital data collected on a wide international scale, and coming from a layered network of administrations and corporations.

This article is structured as follows: section two presents big data assemblage and its various apparatuses; section three frames the research design and explains the method adopted; section four and its sub-sections

show the results of the analysis. The last section concludes with some remarks about the undertaken work and future perspectives.

2. Big Data Assemblages

Critical data studies (CDS) aim at retracing the contextual and relational processes through which data are constructed. One example is research on algorithms (Gillespie 2014; Kitchin 2017), which have concentrated on how algorithms are generated (Bucher 2012; Geiger 2014), or how they worked within specific domains such as journalism (Anderson 2011), security (Amoore 2006, 2009), or finance (Pasquale 2015). A further example of CDS is research on data curation practices. Diesner (2015) affirms that small pre-analytical decisions concerning data preparation for analysis (for example merging, sorting, cleaning, structuring, data reduction, normalization, etc.) −which are often not given careful attention, and about which there are few "best practices" −can have enormous (often undesired) impact on the results of big data research. Finally, some CDS research aims at specifying how cultural, symbolic, and normative values may play a role in promoting certain images of the social world through data. Their objective is the analysis of the connections between the material sphere (technologies, devices, infrastructures) and the socio-cultural one (values, symbols, expert knowledge, disciplinary "discourses", interests, and logic of action). For example, Taylor et al. (2014) demonstrated that the access of corporate big data is proprietary, and that may limit the replicability of studies.

All these pieces of research have focused their analysis on the data assemblage that is "a complex socio-technical system composed of many apparatuses and elements that are thoroughly entwined, whose central concern is the production of a data" (Kitchin and Lauriault 2014, 6). The diffusion of the term assemblage, in French *agencement*, is attributed to the French philosopher Deleuze. He believed that assemblages are entrusted with the function of dismissing the representative thought that arrogates the control of meta-discursive knowledge, of disciplinary specialisms and related institutions. Assemblage is above all the attitude to recognize the production of data as fields of force, heterogeneity of the processes, unforeseen connections in which they are located, and which contributes to produce (Deleuze and Guattari 1980). And just as data are a product of the assemblage, the assemblage is structured and managed to produce those data (Ribes and Jackson 2013). Data and their assemblage are thus mutually constituted. Importantly, they are responsive, dynamic and lively, constantly reconfigured as new data are generated and datasets are combined in different ways (Andrejevic 2013; Beer 2013). Moreover, each data assemblage forms part of a wider datascape (Berry 2011), which encloses the whole spectrum of existing data sources (official statistics, big data on the internet, administrative open data, etc.) and data infra-

structures (data holding, data archives, repositories, etc.) on a specific subject (Aragona and De Rosa 2018). The datascape is therefore composed of many others inter-related and interacting data assemblages and systems.

The fact that any big data assemblage is inextricably linked with other data assemblages makes it hard to empirically isolate it. We have therefore decided to run our research in three European centres of calculation, which produce, use and share digital data. According to Latour (1987), centres of calculation are venues where knowledge production builds upon the mobilization of human (directors, researchers, collaborators, etc.) as well as non-human (documents, books, data, instruments, machines, methods, etc.) resources. He stated that the non-human resources mobilized within centres of calculation by the scientists fulfil three conditions: firstly, they have to be mobile, so they can be transported to a 'centre of calculation'; secondly, they have to be stable to be processed; and thirdly, they have to be combinable in order to be aggregated, transformed and connected to other resources in the process of knowledge production. These properties configure non-human resources as immutable mobiles (*ivi*, 223). Neresini (2015) affirms that digital data have all these three characteristics. They can be shared easily through data infrastructures and digital devices, condensed in numbers and other signs that "are able to communicate meanings that are not direct manifestations of *hic et nunc* subjectivity" (Berger and Luckmann 1966, 58), and finally aggregated, shuffled combined, merged and linked within databases. Data are seen as boundary objects (Star 1989), objects that have different meanings in different social groups, but their structure is sufficiently common to make them acknowledged means of translation. Different from symbols – for every symbol we have a set of stereotypical meaning – the meaning of boundary objects does not come from familiar uses, but is brought to it by the actors who are using and interpreting it in their interaction. Nevertheless, data do not only participate to the formation of knowledge in a symbolic way, but also in a denotative way, giving an active contribution to its construction. As Gitelman and Jackson (2013) argue, data are both framed – actively produced in specific contexts – and framing – themselves producing objects and subjects of knowledge. For example, classifications in social sciences, when acted within institutions, change the ways individuals understand themselves (Hacking 1999). A clear illustration of that is gender statistics, that is the segmentation of any statistical indicator in two categories, men and women. Some kinds of gender inequalities, such as gender pay gap and work-family balance, were not so pressing in society as far as they were measured. Gender statistics helped to claim equality of income between women and men, and a better work-family balance. At the same time, LGBT movements, which defend multiple different gender identities, consider as a discrimination the segmentation of statistics in simply men and women. The problem is that once stabilized, data become autonomous, independent from their construc-

tion procedures and without memory on their origins (Neresini 2015). Data have their own agency, not only because, as symbols, they are cultural products but overall because their meaning, and what they represent, is the result of choices made by a long chain of actors.

Big data assemblages are the joint product of different apparatuses and many competing communities of actors. The apparatuses interact with and shape each other through a contingent and complex web of multifaceted relations (fig.1), with the result of being 'black boxed'.

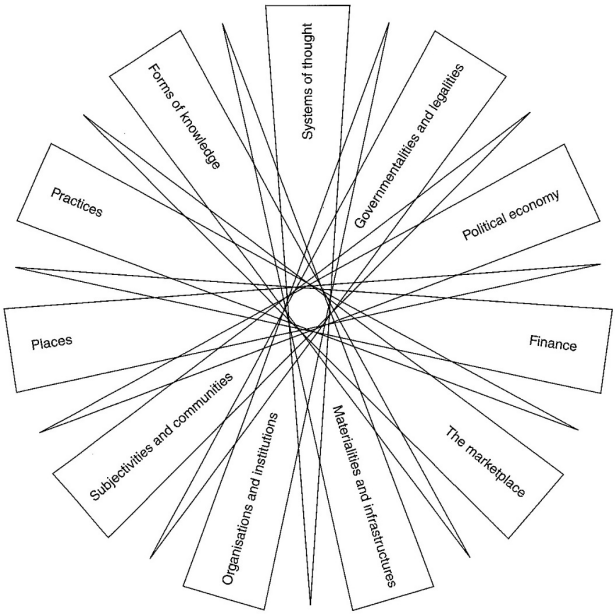

Figure 1 – Apparatuses of the data assemblage. Source: Kitchin (2014b, 26)

In cybernetics, a part of a machine is said black boxed when only the inputs and outputs are known, but not what is in-between. Pasquale (2015) notes that the black box question has been a problem for data even before the advent of big data, because data, whatever its size, are part of different layered activities. It is therefore crucial to follow these elaboration and exchange processes and to retrace the chain of human and non-human actors that compose the big data assemblage. This consists of more than the centre of calculation itself, to include all the technological, political, social and economic apparatuses that frame data. Whenever black boxes are opened, the elaboration processes are revealed, working groups, decisions, competitions and controversies come up (Latour 1987).

Our analysis concentrates on the methodological activities run in the assemblage and on subjectivities and communities. Methodological activities concern techniques, ways of doing, learned behaviours and scientific conventions. They are all the procedural aspects of data, which have changed dramatically in big data assemblages and that mainly refer to the following aspects:

- Data collection: data selection, archive integration techniques, metadata, etc.;
- Data management and organization: responsibilities for data management, intellectual property, consent and ethics, etc.;
- Data analysis: pre-analytics, data mining, text mining, etc.;

Subjectivities and communities refer instead to the different agencies involved in big data assemblage (producers, social scientists, users, etc.) and recall its social aspects. In big data assemblages a dialogue between different kinds of expertise is needed (i.e. statisticians, IT experts, domain experts etc.). Along with this, the socio-technical aspects of data assemblage refer also to the different stakeholders the data are directed to (policy makers, researchers, communication experts, data journalist, citizens). Our analysis focuses on composition of teams (professional profiles, skills, etc.), and the links between the internal actors of the assemblage and other external actors (brokers, corporations, public agencies, etc.). Because it is not possible to separate the apparatuses of the assemblage, by studying methodological activities and subjectivities and communities we have inevitably addressed some questions that are connected with the other apparatuses.

3. Method

Qualitative methods seem suited to deconstruct the contingent and relational nature of big data. We conducted our research on data centres because they are venues where all the apparatuses of data assemblage take form (Aragona et al. 2018). We selected three centres in Europe: Web Science Institute (WSI), Italian National Institute of Statistics (ISTAT) and Norwegian Centre for Research Data (NSD). We chose these three data centres because they are all involved in big data assemblage. They have specific priorities and aims, and different organizational structures; these centres rest also in three different territorial contexts. ISTAT serves the Italian community to produce and communicate official statistics. It is composed of various departments, sections and units that depend from a central executive body. NSD is the Norwegian national archive, and its mission is to help in finding data, and to ensure and control their quality. It has an organizational structure less hierarchical than ISTAT, which is

divided only into three sections (information technology, data services, data protection). Finally, WSI is a research institute within the University of Southampton that has a flat organizational structure without levels of middle management. It aims to undertake interdisciplinary research and to provide insight and intelligence that can lead policy, business strategy, civic engagement and individual choices to meet the social and technical challenges posed by web technologies. These three centres have some common traits that entitled us to compare their activities. At the same time, they have also different characteristics, which allowed us to explore a much wider spectrum of existing sources and of scheme of actors, roles and systems of influence (Aragona and De Rosa 2018; Aragona et al. 2018).

The analysis of data assemblages is usually realized through ethnographies (Geiger 2017; Seaver 2017), we preferred to adopt only qualitative interviewing (in-depth interview and focus group)[1]. The reason for this choice is that we gave priority to the meanings and the relevance that actors participating in the assemblage attribute to the activities they run, according to their role, background knowledge and the context. We run in-depth interviews with directors (2) and heads of sections (7) to encourage a critical reflection on the apparatuses, and a reconstruction of the whole data assemblage. In addition to interviews, we conducted three focus groups – one for each centre – with data team members without managerial functions. Focus groups participants had different educational and professional backgrounds (computer scientists, social and political scientists, statisticians and legal experts on data protection). Focus groups helped us to collect a wider range of opinions, and to explore different procedures. Moreover, they allow us to grasp the relational dynamics between different communities of experts, and their level of engagements in the layered stages of the assemblage.

4. Results

The results of the analysis may be organized in five sections that cover the main problems and challenges that emerged from both the interviews and the focus groups. The first three concerns the methodological aspects (access, selection and interoperability), while the last two focus on the skills needed in the assemblages, and the ethical implications of big data research.

4.1 Access

In recent years, open data initiatives and the building of new data archives and data infrastructures have encouraged the sharing and use of

public data for research. Nevertheless, the problem of data access is still urgent, especially for data produced by private companies:

> When you think of Twitter...the process is massively irritating…it is actually almost impossible to get some kind of data that you want without a special relationship with Twitter. (L., WSI)

According to boyd and Crawford, access may be actually granted to somebody according to their influence, budget and goals: "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):

> Compared to other social networks, we did not used Facebook due the difficulty to access in terms of economic resources; we have used Twitter because it is free. (B., ISTAT)

Mobile-phones operators, app developers, social media providers, retail chain, and surveillance and security companies are under no obligation to share data. Access is therefore usually individually negotiated, and it involves layered networks of agencies and the signs of a series of agreements concerning intellectual property, non-disclosure and re-sharing:

> You can image the effort to get call detailed records; agreements between institutions and authorities, and all kinds of guarantees […]. I spent two years trying to obtain contacts, appointments and agreements. (B., ISTAT)

The question of the access of private data is not a trivial one, because it completely changes the way of thinking about data and their value. When talking about the call detailed records of telephone companies, an interviewee of ISTAT highlights this problem:

> We have never paid for data and we do not want to create a precedent because in my opinion these data are public good; they are not a private property, we all have generated the millions of data by telephoning and they are stored by companies. (A., ISTAT)

This is a clear example of how values come into the activities of the assemblage. In a public data centre, such as ISTAT, data are considered as a public good. The value of public goods is inverse to their scarcity; more the good is diffused, more its value is. On the contrary, in the private market it is scarcity that gives value to goods; rarer is the good, more its value is. For example, a WSI interviewee explains that the access to social data is often constrained and requires agreements with data brokers, specific companies (data aggregators, consolidators and resellers) that allow to buy a large amount of data and layers of services:

> We have got a range of channels for getting social data…One of this to get through is paying intermediate company that gathers social data and provide some added value analysis. (S., WSI)

These findings support the idea that data, are not neutral, impartial expressions of knowledge, but they construct and implement regimes of knowledge (Campbell and Pedersen 2011). Furthermore, they show that the number of intermediaries between the producers and the users of data is growing in big data assemblages. The relations of the centres with data brokers and governmental authorities are just some examples of the multiple possible configurations that the wider networks between the different public and private agencies that participate in the big data assemblage may take.

4.2 Selection

The selection of data emerges in different means. Firstly, interviewees discuss the criteria that orient the choice of big data. Actually, despite the often made claim that big data provides total populations ending our reliance on samples, this is rarely the case for social media data (Highfield et al. 2013). When using data coming from the web, part of the population may not be accessible, because not accessing the internet, or because individuals are passive consumers of internet information, rather than active participants on the web. Respondents wonder about the quality of these data in respect to the selection of a sample from the right population and its representativeness:

> Our purpose is to estimate matrix of flow inter-municipal within both region and province (…) This data source entails methodological problems due to single market share of Telecom[2], then the fact that the same subject can possess more Sim cards and it is not sure that the account holder coincides with who effectively use the Sim card. (C., ISTAT)

Selection errors may become more acute in the case of social media data, because it is more difficult to identify the people and their characteristics:

> The problem would be the quality because social media data are a kind of new data on who are the people. Are the people on Facebook really people? And who are those? The gender, the attitude, the quality of the data comes across that. And that would be one of the main problems. (E., NSD)

> These data as an output of activity in social media are self-selected; you are only analysing people who use Facebook. Twitter is the same and people who use Twitter, although they have a very variable profile and features

> and personalities, there is a common threat and it is that they are Twitter users, and for being a Twitter user you need to have certain treats. That happens with me when I analyse Mooc data as well, to start with I am only analysing learners who are using Moocs. (I., WSI)

Other selection problems are generated by the "velocity and ever-changing nature of big data" that requires a modernization of the organization and of the technologies:

> The structure of the website is always being changed over time and we have to keep up with the technological changes. (…) The velocity and ever-changing nature of big data generates acquisition problems, and it needs the development of new data capturing practices (…) Regards to web scraping (…) it implies a new form of organization than we did until now. It needs to figure out how select the data. (B., ISTAT)

Some critics consider that because the web is changing fast it could make no sense to snapshot phenomena when they can variate very quickly (Lieberman 2008). It is almost impossible to draw any kind of generalization. Selection criticalities are not only technical, but they also require a "new form of organization" able to work with the ever changing form of big data which – as will clearly emerge in section 4.4 on skills – necessitate an overall restructuring of the routinized working activities inside the data centres.

4.3 Interoperability

One of the claims about big data revolution is the possibility to create datasets with strong relationality, which can be combined to generate additional insight and value (Mayer-Schonberger and Cukier 2012). The question of interoperability is not new, and it has been pursued for long time by data infrastructures such as archives, informative systems and repositories. For data to be integrated into new datasets they require shared indexical fields and data standards, consistent metadata systems and compatible format. A broader set of managing and handling problems arises not only with big digital traces on the internet, but also with big data operating in context alongside traditional forms of data, the scaled-up data, what we call "the data that are getting bigger" (Aragona 2016). Data that are getting bigger are research and administrative data that have been integrated, merged, linked and restructured within data infrastructures (i.e. datawarehouses, dashboards, archives, etc.). These have been also named 'small big data' (Gray et al. 2015). It is not always easy to scale-up databases coming from different institutions, because they may be structured on dissimilar standards:

> Well, the data sources that I use have been 2000-3000 institutional repositories around the globe...they attain to specific shared information

> and so from universities all round the world...The problem with that is that it becomes very costly to keep helpful at having an infrastructure which is made by 3000 repositories and 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 have got the different archival and librarian practices in every institution and they will use the software differently and use the metadata alternatively. (L., WSI)

This simultaneous use of various standards calls for metadata harmonization. Metadata standards do not always meet the needs of interoperability between independent standardization communities. The combinations of different specifications seem a core issue for web-based metadata. An interviewee faces the coexistence between multitude of metadata standards with different characteristics:

> We have been using various metadata…Ddi, for instance….Sdmx (…) But what we are working now with is much more on how we can integrate the metadata (A., NSD)

Metadata perform a double function. They help data to become mobile immutables (Latour 1987). The anchoring of data to specific classifications, methods of data collections and procedures keeps them stable, as well as increases their mobility, because it eases the combination with other data. At the same time, metadata facilitate the development of standardized procedures for the management of data flows, which may be implemented in different data assemblages. For example, The Generic Statistical Business Process Model (GSBPM) introduces a new methodology to connect traditional research (survey and administrative data) with big data within National Statistical Institutes, integrating data and metadata standards and harmonizing statistical computing processes[3]:

> GSBPM has spread starting from Unece that it introduced this methodology to standardize the process within National Statistical Institutes. We are trying to introduce and connect the production with big data in this scheme (GSBPM) to represent and standardize each modification on traditional flow of the model for the purpose of replicability and transparency. (B., ISTAT)

Metadata specifications and standard processes therefore add further value that may enhance the combination, exchange and reuse of (big) data coming from different sources. Examples are the data stored by social science data archives such as Cessda, the Central European Social Science Data Archive, or the Information Systems that have been built by Eurostat and National Statistical Institutes. One problem is to handle these data to prospective users.

> And then we have the problem of storing and organising what we have produced and get access to. We have all kinds of metadata problems, how you describe a document, a data, a service so that is easy to find? And then we have the (…) discovery and dissemination systems: how do you push out the data again to the prospective users? And how do we make them able to analyse the data? What kind of statistical packages are they using? Are they using Salstat, Spss or whatever could be…How do you create flexibility? There is a big difference in data format if you want to use Spss versus R or Stda to do the analytic work. (A., NSD)

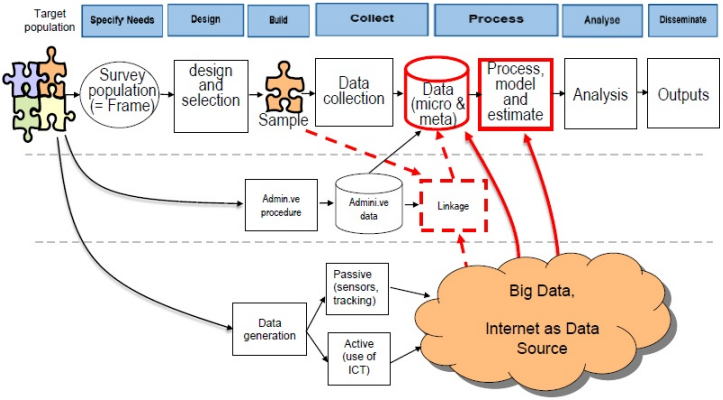

Figure 2 – The use of Big Data in Italian Official statistic according to Generic Statistic Business Process Model. Source: De Francisci (2017)

The adoption of common standards may offer more complete documentation, more widespread know-how and better access to reusable tools. Unlike traditional data assemblages that stopped when data were released, current big data assemblages must follow up on the way data are handled to final users through platforms, infrastructures and the media. Interviewees wonder if users are able to transform these data into knowledge, and how this process works (Giovannini 2014).

4.4 Skills and communities

The lack of the proper skills for handling big data in statistical offices is a challenge that needs to be addressed (Baldacci 2016). Indeed, statisticians, experts of fields, computer scientists, and all the other communities of experts who for long have been dealing with data are significantly affected by the assemblage of big data:

> We have been tackling new problems: for now, problems of production…we had response problems…so different troubles than we were used to face, however I would say that we are equipping ourselves with a new instrument and acquiring a culture which is in line with that is happening in the world, such as data science, technics of machine learning, production of models…we are approaching these tools and using them jointly the methodological tools that we already have. (B., ISTAT)

The skills needed are not simply technical, but also deeply epistemological, which consist of the ability of mixing social theory and computation, data and modelling in an innovative way. In this respect, the same interviewee continues reporting the lack of big data experts with these skills on the labour market, and he affirms that the higher education is not sufficiently focused on targeting big data:

> I think that university should provide more competences to the students to work with these kinds of data (…) only in the last years they have started to set up master focused on data. But in the next few years, we expect a major demand on the labour market (…) machine learning, the skills about the statistic but also the new skills relating to data science. (B., ISTAT)

Apart from the new skills, in big data assemblages a dialogue between the different communities of experts is required to blend methodologies and disciplinary matrixes, and shape what Lakatos (1976) called background knowledge (the whole set of facts and parameters used in the construction of any given theory, and of any given data):

> I have been lucky enough to come across and work with people in all their disciplines that have not been to heavily shade by their own discipline which means they are still "malleable" and this means the way you approach a problem, and the universe from which you depart it is negotiable and is negotiated. We have been able to easily accept that there are other ways to see the world and other ways to get to a conclusion or other ways even to name it. (L., WSI)

One interviewee traces a distinction between interdisciplinarity and multidisciplinarity. The former is supposed to be a new thing that comes out from the blending of concepts and backgrounds from different disciplines. It is related to the overcome of some political struggles between scientific communities. The latter is just limited to the sum of the different concepts and methods borrowed from the various branches of research and knowledge:

> Multidisciplinarity is very rich and very useful, but interdisciplinary is far beyond it, because it demands that if you have some stand points and others contribute, if there is synergy between them they can make-up with

> an ordered new think and this is interdisciplinarity about. It is not only collaboration between two disciplines, it is to come up with something new that all of them can agree and can transport on their research field. (…) For me, multidisciplinarity, I agree, we can talk with people from different discipline sets, we share knowledge, and it's very useful, but interdisciplinarity is more than this, that's my point. (P., WSI)

Therefore, an interdisciplinary context fosters the discussion between experts and greater openness in approaching a problem. As Berger and Luckmann noted when talking about the maintenance of symbolic universes, a pluralist situation mines the capacity of the definition of reality based on traditional symbolic universes and of resisting to changes. Pluralism: "encourages skepticism and innovation, it is intrinsically subversive of *status quo* taken for granted reality" (1966, 174). According to the interviewees, a pluralism of disciplines seems to be a key aspect of transition from data to big data assemblages.

4.5 Ethics

According to the EU Parliament[4], European citizens should become aware not only of their digital rights, but also about algorithmic governance, automated data processing, and means of collecting data (web scraping, social networks, etc.). Yet the differences in the legal frameworks, and high bureaucratization have been obstacles for research collaboration and data sharing across national borders. For this reason, European Union adopted the regulation on personal data protection, the *General Data Protection Regulation* (GDPR), to safeguard the privacy of EU citizens. GDPR regulates data breaches notification, right to access, right to be forgotten and data portability. It pursues the creation a common legal framework that can push cross-national research through trust common legislation and harmonized practices. This new legislation should guarantee the rights and privacy of the citizens, fostering a greater control on their own data:

> The main reasons on the process of making the GDPR started are all these new fonts of data and all the data a lot of people do not know their rights, they do not have control over their data (…) And move from regulations that shows it will be implemented more or less in the same way in all the European countries. (S., NSD)

Big data seem to challenge the entire ethical system that has been created and institutionalized on different kinds of data:

> I think that the kind of data that existed has shaped out the structure of the ethical regulation system (…) But I think that the new form of data that we have challenges the ethical system that we have as a bureaucratised system. (S., WSI)

Nevertheless, an interviewee explains that GDPR does not fully overcome the problems of big data research, rather it has a limited flexibility in its application. Specifically, the access to data is still costly and time consuming; each authority requires information about the project with descriptions and justifications for the processing of personal data:

> That is one of the many issues for big data researchers in any industry or GDPR is the data limitation. One of the main aims of a data researcher is that it should collect all the data that you can gather and see if you can find a pattern. So, I think that the GDPR and big data researchers are difficult to combine. (M., NSD)

The ethical concerns are more urgent with social media data. As the case of *Cambridge Analytica* has shown, mapping personality traits based on what people had liked on Facebook, and then use that information for profiling and influencing citizens may rise important ethical implications because, as an interviewee notes, it is somewhat obscure who these data can be handled to:

> I think, we are quite good with research ethics… but, I think, as we generate more and more data with social media, in particular, when you look at the terms and conditions of things like Instagram or Facebook, we are really lowering the expectation bar of how people treat data and use data. Who you can give it to? What you can use it for? (L., WSI)

These findings show that new ethical regulations may reinforce hyper-networked ethics. Floridi (2013) refers to this as "infra-ethics", where at least three main stakeholders are affected: data generators, data collectors and data users. Alike our interviewees, the agents in this network may have different opinions about data ethics. Since they interact with other actors within the data assemblage, they may cause collateral consequences on all the others by facilitating or hindering ethic actions.

5. Conclusions

The research shows that data happen through structured social practices "in and through which various agents and their interests generate forms of expertise, interpretation, concepts, and methods" (Ruppert et al. 2017, 3). By inspecting the work of data centres of calculation, we were able to identify some stabilized activities (for example the establishment of agreements at different degrees of formality with data providers and data brokers) and to assess their consequences on data quality (for example on representativeness). In addition, we addressed the effects of some criticalities on the whole big data assemblage. One example is the lack of

interoperability, which can affect the timeliness and accessibility of big data. Furthermore, we retraced the different communities of experts that participate to the processes of the assemblage. Big data assemblages are imbued with multidisciplinarity. On one side, this is needed because big data requires multiple computational, statistics and domain expertise. On the other side, pluralism of disciplines is seen as a way to improve adaptability and enhance innovation. Finally, the specific layered activities of big data assemblage are throughout concerned with ethics, but they all pose various ethical problems to be overcome, and a size fits all solution does not emerge from the interviews.

The analysis brings some valuable insights about the problematic issues related to big data assemblages. A central question is how we could arrive at better conventions that can help an effective use of big data. Access constraints, acquisition problems, selection biases and pre-analytical work may be problematic unless a series of routinized activities takes place. Conventions are necessary to fix standards that insure the quality of data, and in our opinion, an institutional setup – as the one is moving its first step forward inside ISTAT and the others European statistical institutes – is a very reasonable thing to wish for. This institutional setup has served so well in the case of survey data, for example through the standard definition of the total survey error, the adoption of classification standards and the exchange of metadata. The establishment of routinized activities is strictly connected with the experts needed inside big data assemblage. The lack of skills lamented by the more established data centres may hinder the development of big data assemblages and their effective functioning. Moreover, the ever-changing nature of big data infrastructures, platforms and interfaces involves not only acquiring new skills from outside the centres, but also constantly, and probably costly, updating the expertise and capacities required to run the activities of big data assemblages.

The methodological posture adopted in this paper allowed us to pick up choices, compromises and agreements and to unveil black boxed aspects of big data assemblage. The comparative focus on the three centres of calculation entailed us to disentangle the different resources (human and non-human) mobilized within the assemblages and to explore "from below" – through the words of the main actors participating in the assemblage – the contingent and contextual making of big data. This piece of research stresses that the definition of data and big data should be always seen as a product of a convention and subjected to debate. By isolating and inspecting some methodological aspects of the big data assemblage (i.e. access, selection and interoperability) it is possible to increase the awareness that data are not given, but actively constructed through socio-technical practices.

Our study should be seen as an attempt to grasp the complex apparatuses that form big data assemblages, because it concentrates only on the socio-technical practices of big data production and management,

and it is confined to the study of big data in social research. Further work should isolate some applications of big data (i.e. government or business) in order to observe how they are brought to use within different communities of stakeholders and users, and to reconstruct the practices within the other apparatuses of the assemblage.

References

Alonso, W. and Starr, P. (eds.) (1987) *The Politics of Numbers*, New York, Russell Sage Foundation.

Amoore, L. (2006) *Biometric Borders: Governing Mobilities in the War on Terror*, in "Political Geography", 25, pp. 336-351.

Amoore, L. (2009) *Algorithmic War: Everyday Geographies of the War on Terror*, in "Antipode", 41, pp. 49-69.

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

Anderson, C. W. (2011) *Deliberative, Agonistic, and Algorithmic Audiences: Journalism's Vision of Its Public in an Age of Audience Transparency*, in "International Journal of Communication", 5, pp. 529-547.

Andrejevic, M. (2013) *Infoglut: How Too Much Information Is Changing the Way We Think and Know*, New York, Routledge.

Aragona, B. (2016) *Big data o data that are getting bigger?*, in "Sociologia e Ricerca Sociale", 109, pp. 42-53.

Aragona, B. and De Rosa, R. (2018) *Policy Making at the Time of Big Data: Datascape, Datasphere, Data Culture*, in "Sociologia Italiana", 11, pp. 173-187.

Aragona, B., Felaco, C. and Marino, M. (2018) *The Politics of Big Data Assemblages*, in "PArtecipazione e COnflitto", 11 (2), pp. 448-471.

Baldacci, E. (2016) *Innovation in Statistical Processes and Products: A European view*, Data Science and Social Reserch, Naples, Febraury 17-18.

Beer, D. (2013) *Popular Culture and New Media: The Politics of Circulation*, Basingstoke, Palgrave Macmillan.

Berger, P. and Luckmann, T. (1966) *The Social Construction of Reality*, New York, Double and Company.

Berry, D. M. (2011) *The Computational Turn: Thinking about the Digital Humanities*, in "Culture Machine", 12, http://sro.sussex.ac.uk/49813/ (Retrieved April 25, 2019).

boyd, D. and Crawford, K. (2012) *Critical Questions for Big Data*, in "Information, Communication and Society", 15 (5), pp. 662-79.

Bucher, T. (2012) *'Want to be on the top?' Algorithmic Power and the Threat of*

Invisibility on Facebook, in "New Media and Society", 14 (7), pp. 1164–1180.

Campbell, J.L. and Pedersen, O.K. (2011) *Knowledge Regimes and Comparative Political Economy*, in "Ideas and politics in social science research", 167, pp. 172-90.

Dalton, C. and Thatcher J. (2014) *What Does a Critical Data Studies Look Like, and Why Do We Care? Seven Points for a Critical Approach to 'Big Data'*, in "Society and Space open site", https://societyandspace.org/2014/05/12/what-does-a-critical-data-studies-look-like-and-why-do-we-care-craig-dalton-and-jim-thatcher/ (Retrieved April 25, 2019).

De Francisci, S. (2017) *Big Data e Linked Open Data per la statistica ufficiale*, Forum PA 2017, May, 25.

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

Desrosières, A. (2010) *La politique des grands nombres: histoire de la raison statistique*, Paris, Editions La Découverte.

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

Diesner, J. (2015) *Small Decisions with Big Impact on Data Analytics*, in "Big Data & Society", 2 (2).

Feyerabend, P. (1969) *Science without Experience*, in "Journal of Philosophy", 56, pp. 791-794.

Floridi, L. (2010) *Information: A Very Short Introduction*, Oxford, Oxford University Press.

Floridi, L. (2013) *The Philosophy of Information*, Oxford, Oxford University Press.

Geiger, R. S. (2014) *Bots, Bespoke, Code and the Materiality of Software Platforms*, in "Information, Communication & Society", 17 (3), pp. 342-356.

Geiger, R. S. (2017) *Beyond Opening Up the Black Box: Investigating the Role of Algorithmic Systems in Wikipedian Organizational Culture*, in "Big Data & Society", 4(2).

Gillespie, T. (2014) *Algorithm [draft] [#digitalkeyword]*, in "Culture Digitally", http://culturedigitally.org/2014/06/algorithm-draft-digitalkeyword/ (Retrieved April 25, 2019).

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.

Gitelman, L. and Jackson, V. (2013) *Introduction*, in L. Gitelman (ed.) *"Raw data" is an oxymoron*, Cambridge, MIT Press, pp. 1-14.

Gray, E., Jennings, W., Farrall, S. and Hay, C. (2015) *Small Big Data: Using Multiple Data-Sets to Explore Unfolding Social and Economic Change*, in "Big

Data & Society", 2 (1).

Hacking, I. (1999) *The Social Construction of What?*, Harvard, Harvard University Press.

Highfield, T, Harrington, S. and Bruns, A. (2013) *Twitter as a Technology for Audiencing and Fandom*, in "Information, Communication and Society", 16(3), pp. 315-339.

Iliadis A. and Russo, F. (2016) *Critical Data Studies: An Introduction*, in "Big Data & Society", 3 (2).

Kitchin, R. (2014a) *Big Data, New Epistemologies and Paradigm Shifts*, in "Big Data & Society", 1 (1), pp.1-12.

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

Kitchin, R. (2017) *Thinking Critically About and Researching Algorithms*, in "Information, Communication & Society", 20 (1), pp. 14-29.

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

Kuhn, T. S. (1962) *The Structure of Scientific Revolutions*, Chicago, University of Chicago press.

Lakatos, I. (1976) *Proof and Refutations*, in J. Worrall and E. Zahar (eds.), *The Logic of Mathematical Discovery*, Cambridge, Cambridge University Press.

Latour, B. (1987) *Science in Action*, Cambridge, Harvard University Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., Cristakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Van Alstyne, M. (2009) *Computational Social Science*, in "Science", 323 (5915), pp. 721-723.

Leonelli, S. (2016) *Data-centric Biology: A Philosophical Study*, Chicago, University of Chicago Press.

Lieberman, D. Z. (2008) *Evaluation of the Stability and Validity of Participant Samples Recruited over the Internet*, in "CyberPsychology & Behavior", 11 (6), pp. 743-745.

Manovich, L. (2001) *The Language of New Media*, Cambridge (MA), MIT Press.

Mayer-Schönberger, V. and Cukier, K. (2012) *Big Data: A Revolution that Transforms How We Work, Live, and Think*, Boston, Houghton Mifflin Harcourt Boston.

Neresini, F. (2015) *Quando i numeri diventano grandi: che cosa possiamo imparare dalla scienza*, in "Rassegna Italiana di Sociologia", 56 (3-4), pp. 405-431.

Pasquale, F. (2015) *The Black Box Society: The Secret Algorithms that Control Money and Information*, Harvard, Harvard University Press.

Ribes D. and Jackson S. J. (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 Bigo, D. (2017) *Data Politics*, in "Big Data & Society": 4 (2), pp. 1-7.

Salais, R. and Storper, M. (1993) *Les mondes de production*, Paris, Editions del'EHESS.

Seaver, N. (2017) *Algorithms as Culture: Some Tactics for the Ethnography of Algorithmic Systems*, in "Big Data & Society", 4 (2).

Star, S. L. (1989) *Regions of the Mind: Brain Research and the Quest for Scientific Certainty*, Stanford University Press.

Taylor, L., Schroeder, R. and Meyer, E. (2014) *Emerging Practices and Perspectives on Big Data Analysis in Economics: Bigger and Better or More of the Same?* in "Big Data & Society", 1 (2), pp. 7-16.

Thévenot, L. (1984) *Rules and Implements: Investment in Forms*, in "Information, International Social Science Council", 23 (1), pp. 1-45.

[1] The interviews in ISTAT were conducted in Italian language and then translated, while those conducted in WSI and NSD were transcribed verbatim.

[2] Telecom is an Italian telecommunication company that offers fixed and mobile communication services.

[3] For details, see https://statswiki.unece.org/display/GSBPM/I._Introduction #I._-Introduction-_Toc375051192.

[4] For details, see http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0076+0+DOC+XML+V0//EN.