Citizen Data and Trust in Official Statistics

Evelyn Ruppert*, Francisca Grommé,* Funda Ustek-Spilda and Baki Cakici*****

Abstract – From smartphones, meters, fridges and cars to internet platforms, the data of digital technologies are the data of citizens. In addition to raising political and ethical issues of privacy, confidentiality and data protection, this calls for rethinking relations to citizens in the production of data for statistics if they are to be trusted by citizens. We outline an approach that involves co-producing data with citizens as partners of statistical production, from the design of a data production platform to the interpretation and analysis of data. While raising issues such as data quality and reliability, we argue co-production can potentially mitigate problems associated with the re-purposing of Big Data. We argue that in a time of "alternative facts", what constitutes legitimate knowledge and expertise are major political sites of contention and struggle and require going beyond defending existing practices towards inventing new ones. In this context, we contend that the future of official statistics not only depends on inventing new data sources and methods but also mobilising the possibilities of digital technologies to establish new relations with citizens.

JEL Classification: A14, C93, O35, O38
Keywords: citizen science, co-production, experimentalism, privacy-by-design, smart statistics

** Department of Sociology, Goldsmiths University of London (E.Ruppert@gold.ac.uk; F.Gromme@gold.ac.uk)*
*** Department of Media and Communications, London School of Economics (f.ustek-spilda@lse.ac.uk)*
**** Technologies in Practice, IT University of Copenhagen (bakc@itu.dk)*

We would like to thank two anonymous reviewers for their comments and suggestions. We are also grateful for the support and involvement of numerous national and international statisticians who made this research possible at the UK Office for National Statistics, Statistics Netherlands, Statistics Estonia, Turkish Statistical Institute, Statistics Finland, Eurostat and UNECE. Funding from a European Research Council (ERC, Consolidator Grant – Agreement n° 615588) supported the writing of this article. Principal Investigator, Evelyn Ruppert, Goldsmiths, University of London.

To cite this article: Ruppert, E., Grommé, F., Ustek-Spilda, F. & Cakici, B. (2018). Citizen Data and Trust in Official Statistics. *Economie et Statistique / Economics and Statistics*, 505-506, 179–193. https://doi.org/10.24187/ecostat.2018.505d.1971

National Statistical Institutes (NSI) experiments concerning the potential of Big Data generated by various digital technologies as a new source for the making of official statistics have now been underway for about five years. These have led to the identification of several concerns such as data access, data ownership, privacy and ethics, data representativeness, data quality and so on. Amongst other things, these concerns are understood as potential risks to the reputation and public image of NSIs working with Big Data sources, as identified in a report of the UNECE Big Data Privacy task team (UNECE, 2014). That report summarised a number of strategies to mitigate such risks including the enforcement of ethical principles through instruments of accountability and informed consent; establishing strong compliance controls; developing monitoring systems to track reputational threats; ensuring transparency and understanding through clear communication with stakeholders about the use of data and the organisation of dialogues with the public; and creating a crisis communication plan. The report also argued, as have others produced by international bodies such as Eurostat's Big Data Task Force, that repurposing Big Data sources not only presents technical challenges but potentially could undermine citizens' trust in how NSIs generate data and produce official statistics. Similar challenges are encountered when NSIs seek to repurpose administrative data generated by other government departments, which has introduced not only technical challenges but for some NSIs also raised concerns about how data is shared, joined up and used for purposes other than for what they were originally generated.

Of course, questions of citizen trust in official statistics are not new. While trust is also a concern in relation to other stakeholders including ministries, government agencies, media, universities, and other public or private research organisations that rely on official statistics, it is trust in relation to citizens that concerns us here. The history of established methods of generating social and population statistics, such as census questionnaires, surveys and time diaries demonstrates that elaborate practices have been required to secure citizens' trust in how data is generated and used for official statistics. Through practices such as focus groups, the pilot testing of questions, and consultations with civic organisations about issues of consent, data protection, privacy, impartiality and professional standards, NSIs have sought to secure the trust of citizens (Struijs *et al.*, 2014, p. 2). Understood in this way, trust is not the result of one but myriad practices through which the trustworthiness of official statistics is accomplished.

Big Data, because it is generated not by governments but private corporations such as platform owners, if used for official statistics could undermine these practices and the trust they have relatively well performed. As some statisticians have noted, "[of] critical importance is the implication of any use of Big Data for the public perception of a NSI as this has a direct impact on trust in official statistics" (Struijs *et al.*, 2014, p. 3). While Struijs *et al.* argue that such risks can be mitigated by other practices such as "being transparent about what and how Big Data sources are used", we suggest that while necessary this would be insufficient due to another significant issue: the repurposing of Big Data for official statistics constitutes a break and detachment in the relation between NSIs and citizens. While not without problems, established methods such as those noted above have involved more-or-less direct relations between NSIs and citizens to secure data as a collective accomplishment and social good. These relations enable citizens to be relatively active in their identification such as how they translate their knowledge and experiences into responses to questions and, we suggest, in turn contribute to accomplishing trust in and the legitimacy of official statistics.

This proposition was initially put forward in the "Socialising Big Data" project, which involved collaborative workshops with national and international statisticians and led to a proposal for a social framework for Big Data (Ruppert *et al.*, 2015). The framework posited models of social ownership that stress sharing, collaborative, and co-operative possibilities and that imagine Big Data as a social and collective rather than private resource. The approach that we develop in this article builds on this aspiration to develop the concept of "citizen data" as a form of "re-attachment" and social ownership that establishes new relations with citizens as co-producers of data for official statistics rather than as ever more distant subjects whose impressions and confidence need to be managed.

We contend that this understanding of new relations is critical in two ways. First, unlike some uses of the term that define citizen data as data about citizens, our conception recognises that Big Data and citizens are inseparable: the data of digital technologies is the data

of citizens. Second, relations that involve more direct engagements with citizens are necessary to address another consequence of detachment when data such as that generated by social media, mobile phones and browsers is repurposed: the risk of a widening gap between citizens' actions, identifications and experiences and how they are categorised, included and excluded in statistics, the interpretation of that data, and citizens' identifications with the resulting statistics.[1] We refer to this risk as a widening gap because these consequences are not entirely new or limited to Big Data.[2] Former Eurostat Director General Walter Radermacher expressed this more generally as a gap between citizen experiences and official statistics which in turn calls for "subjective statistics".[3] In saying so he stressed the need for a more democratic debate between citizens and data producers and owners to achieve a "more subjective, differentiated understanding of our world", instead of "technocrats and politicians sitting together and confronting citizens in the end".[4] For our concept of citizen data this requires processes of co-production that involve direct relations with citizens in the production of data for making official statistics.

Our argument draws on several years of fieldwork conducted at NSIs and international statistical organisations (see Box; see also the working paper by Grommé *et al.*, 2017). This research led to the identification of four principles for citizen data that started from key "matters of concern" statisticians have expressed about the future of official statistics which we encountered in our fieldwork. We consider these as matters of concern for two reasons. First, to recognise them as normativities that influence and guide statisticians' actions and development of practical solutions (Boltanski & Chiapello, 2007). Second, to engage in a form of critique that does not dismiss the concepts of our research subjects but first engages with how they conceive and define concepts to then consider how concepts can be reconceived (Latour, 2004). That is, taking up the concerns statisticians have expressed does not mean to agree with them and their assumptions but to engage with and then reconceive those concerns. The four matters of concern we identified as significant to our concept of citizen data are experimentalism, citizen science, smart statistics and privacy-by-design. In the next part of this article we introduce each concern and then draw on a range of literature in the social sciences to reconceive each and then express them as principles of citizen data. Central to our

1. For example, experiments with mobile phone data to model mobility encounter problems when attempting to interpret the meaning of travel patterns.

2. We are aware that issues of representation also affect established statistical methods. GDP, Gross Domestic Product, for instance, is one such highly debated official statistic. Columbia University economist Joseph Stiglitz draws attention to how GDP has come to be "fetishised" as "the" indicator of how well a national economy is doing, despite various shortcomings (Stiglitz et al., *2009). Consequently, Fleurbaey (2009) suggests moving "beyond GDP" and draws attention to other approaches, including recent developments in the analysis of sustainability, happiness and the theory of social choice and fair allocation to the studies of social welfare. Similar arguments have also been raised for employment indicators, especially with respect to people working in non-regular employment arrangements (see Hussmanns, 2004).*

3. Fieldwork notes*, Eurostat conference "Towards More Agile Social Statistics", Luxembourg, 28-30 November 2016.*

4. Idem.

Box – **The research project**

Our concept of citizen data comes from several years of ethnographic fieldwork that we conducted at five NSIs and two international statistical organisations, which involved observing conferences and meetings, following and analysing publications, and conducting interviews and engaging in conversations with statisticians. More precisely, this article builds on and summarises key points in an ARITHMUS working paper by Grommé *et al.* (2017). ARITHMUS (Peopling Europe: How data make a people), an ERC funded project, began in 2014 with a team of six researchers: Evelyn Ruppert (Principal Investigator), Baki Cakici, Francisca Grommé, Stephan Scheel, and Funda Ustek-Spilda (Postdoctoral Researchers), and Ville Takala (Doctoral Researcher). We followed working practices at five NSIs (UK Office for National Statistics, Statistics Netherlands, Statistics Estonia, Turkish Statistical Institute, and Statistics Finland) and two international organisations (Eurostat and UNECE). Amongst other things, we followed statisticians' debates about and experiments with digital technologies and big data and their implications for official statistics. Based on this fieldwork we conducted two workshops with a project advisory group of statisticians to discuss some of our analyses such as the changing relations between NSIs and citizens as a consequence of new digital technologies and big data sources. This led to a working paper that summarised some of the arguments outlined in this article and introduced the concept of citizen data, which was reviewed by the advisory group (Grommé *et al.*, 2017). That review led to a collaborative workshop with the advisory group and a broader group of statisticians, academic researchers, information designers and facilitators on the development of design principles for the co-production of an app for citizen data. Rather than summarising empirical material from our ethnography and the workshops, our objective here is to outline the conception of citizen data that we have developed as a result of this research.

re-conception is that the future of official statistics not only depends on working with new digital technologies, data sources and inventing methods, but on establishing new relations to citizens (Ruppert, 2018).

We have intended this discussion of a concept of citizen data principally for statisticians but also for social science researchers for three key reasons. One is that we have brought concepts and understandings advanced in the social sciences to bear on matters of concern expressed by statisticians. In this way, we contribute more generally to social science research methods. Another reason is that the principles and concept of citizen data also apply to debates within the social sciences concerning research methods that engage with digital technologies and Big Data sources. That is, while the issues and objectives of social science research are different, relations to citizens in the production of knowledge are a shared concern. Third, as reflected in our research method which involved workshops with statisticians, a concept of citizen data calls for experimental engagements not only with citizens but also between social scientists and statisticians.

Experimentalism

The first matter of concern that we have come across in our fieldwork is experimentalism. Government agencies and corporations have embraced experimentation as a necessary part of innovation. Official statistics is a good example as attested by the development of innovation laboratories, sandboxes, hackathons and exploratory research projects.[5] For statisticians, experiments with new digital technologies and Big Data are methods to develop new ways of thinking, techniques, and skills in the production of official statistics. There are also various strands within the social sciences that engage with experimentalism. Relatively new, however, is the adoption of experimenting as a method to open scientific and technological expertise to different actors to generate new ways of thinking. In areas as diverse as wheelchair design, Big Data and synthetic biology, social scientists have adopted experimentalism to generate new spaces of problem formulation, engage with different actors and consider different possibilities.[6] That is, a key premise is that experimental modes of *collaboration* can generate new ways of thinking.

Broadly speaking, we can distinguish two models through which collaborative experiments may seek to achieve this. The first is through various forms of participation intended to achieve a degree of democratisation by opening up scientific and technical debates and processes to publics (Marres, 2012). The second is to experiment collaboratively to develop and explore new problem formulations, transcend ingrained styles of reasoning, disrupt existing hierarchies and critically examine how knowledge is created (Rabinow & Bennett, 2012). This is the model of a "collaborator" (or, co-laboratory) where participants engage in the common exploration of a topic. The Socialising Big Data project previously mentioned engaged with this model by conducting workshops and discussions with national statisticians, genomic scientists and waste management engineers to define and develop shared concepts for understanding Big Data (Ruppert *et al.*, 2015). Another form of collaboration involves the co-production of a "thing" – a tangible end-product – through which collaborators *practically* explore and develop shared concepts and issues. Working on a common product makes "issues experimentally available to such an extent that "the possible" becomes tangible, formable, and within reach" (Binder *et al.*, 2015, p. 12). As a method, it forces participants to make future modes of working explicit (Muniesa & Linhardt, 2011). Generally, from the social studies of science we learn that such collaborative experiments also require reshaping relations between participants, technologies and knowledge. This is also a principle of what is called in the social sciences and humanities, practice-based research, which involves an engagement between participants and the skills, materials, small tasks and everyday labour, in addition to texts and spoken word, that are enrolled in making things (Jungnickel, 2017). Making things, as opposed to unravelling or deconstructing them, involves a close entanglement with different participants and can increase understanding of the skills, relations and infrastructures that are part of an end-product (*ibid.*).

Experimentalism is especially recognised as a necessary approach to uncertainty and change. For example, in an article on a collaboration between academics, farmers and environmentalists, Waterton and Tsouvalis (2015, p. 477) ask how "the politics of nature can be envisioned for an age conscious of the complexity,

5. See for example, experimental statistics produced by Eurostat: http://ec.europa.eu/eurostat/web/experimental-statistics/.

6. For these three examples, see: https://entornoalasilla.wordpress.com/english/; Ruppert et al.*, 2015; and http://www.anthropos-lab.net/about.*

contingency, and relationality of the world?" They investigate a collaboration between themselves as social scientists with environmental experts and farmers to improve water quality. In their experience, a shared inquiry opened up questions of how to understand water pollution: in terms of isolated causes or wider socio-technical relations and histories. They thus adopted an agenda of experimentation that understands the generation of knowledge as involving "hybrid forums" (Callon *et al.*, 2011) or "new collectives" (Latour, 2006) in which participants reflexively engage in reconstructing the relations, histories and stakeholders involved in an issue. Uncertainty is not something to be solved, instead it needs to be acknowledged and worked with in an ongoing collective process of knowledge production. In practice this entails a "care-full" approach (Grommé, 2015) which entails the exercise of responsibilities for monitoring and documenting who and what are (unavoidably) included and excluded; avoiding ambiguity about the terms of evaluation by making explicit how outcomes are assessed; recognising that failure is likely caused by myriad factors; and, understanding that values are inseparable from facts. "Care-full" therefore does not only refer to a cautious approach, but also active acknowledgement that experiments continually reshape relations and redistribute effects in sometimes unexpected ways.

As a principle of citizen data, experimentalism thus involves not only experimenting but collaborating to make ways of thinking and generating knowledge "open" to the influence and insights of others and in doing so imagining and speculating on alternatives and possibilities (Stengers, 2010). It requires being accountable to and accounting for the procedures and practices of experiments. Finally, it means being open to how relations between different participants in the making of knowledge might be organised differently. Taking up our point on new relations between citizens and NSIs, experimentalism thus involves active and open forms of participation and influence. We develop this further through a second principle, that of citizen science, to explore how relations between NSIs and citizens in the making of data and official statistics might further be reconceived.

Citizen Science

Some statistical organisations have started experimenting with models of citizen engagement in the production of data. Such models often draw on existing conceptions of citizen science, which we will briefly discuss here to explore how we might reconceive them. Different models of citizen science conceive of citizens as not only research subjects, but as actively involved in the production of data as opposed to traditional methods where they are usually understood as respondents. There are many definitions and interpretations of citizen science and the terms of citizen engagement in the making of data. The European Commission (EC), for example, defines it as the "production of knowledge beyond the scope of professional science, often referred to as lay, local and traditional knowledge" (European Commission, 2013, p. 5). Goodchild (2007) uses the term to describe communities or networks of citizens who act as observers in some domain of science. This is the most commonly accepted definition especially evident in the significant momentum citizen science has gained in the natural sciences in recent years (Kullenberg & Kasperowski, 2016, p. 2). However, the practice of engaging people in collecting and submitting data for scientific purposes goes back at least to the 1960s, though the term itself was not used until the 1990s (*ibid.*).[7]

A second version involves citizens not as only observers but co-producers or producers of scientific studies and data to reflect their own concerns, needs and questions. This version includes local and activist-oriented approaches referred to as "community based auditing", "civic science", "community environmental policing", "street science", "popular epidemiology", "crowd science", and "Do It Yourself Science" (Kullenberg & Kasperowski 2016, p. 2). These versions range from citizens seeking close alliances with scientific and knowledge institutions to citizens engaging in the production of independent knowledge together with scientists.

Citizens' objectives for engaging in scientific data production are multiple, ranging from documenting concerns about environmental issues, to creating online archival maps of local historical sites or transcribing Shakespearean contemporaries.[8] Goodchild (2007, p. 219) suggests that people who generally participate and share information on the internet are more likely to volunteer geographic information and

7. For some researchers, it includes the National Audubon Society's Annual Christmas Bird Count in early 1900s, where citizens participated in the observation and enumeration of bird species.
8. Some of these examples are documented at www.zooniverse.org.

contribute to data collection initiatives such as OpenStreetMap (OSM). On this basis he argues that two kinds of people are likely to participate: people who seek self-promotion and volunteer personal information on the internet to make it "available to friends and relations, irrespective of the fact that it becomes available to all"; and, people who seek personal satisfaction derived from contributing anonymous information and seeing it appear as part of a developing "patchwork" of collective contributions (*ibid.*, p. 219).

Jasanoff (2003) notes that models of citizen science can facilitate meaningful interaction among policymakers, scientific experts, corporate producers and publics (pp. 235–236). She argues that the pressure for accountability in expert decision-making is manifest in the demand for greater transparency and wider participation. However, participatory opportunities cannot alone ensure the representative and democratic governance of science and technology. Jasanoff underscores that the attention of modern states has focused on refining "technologies of hubris" that are designed to facilitate management and control by bracketing off uncertainty, political objections and the unforeseen complexities of everyday life (p. 238). What is lacking is not just knowledge, but ways to bring uncertain, unknown processes and methods into the dynamics of democratic debate (pp. 239–240). For this reason Jasanoff suggests citizen science as a possible model of democratic interaction between different stakeholders in the production of science. In this way citizen science models can be thought of as "technologies of humility", that is, *social* technologies that involve relations between governments, decision-makers, experts, and citizens in the management of technology for "assessing the unknown and the uncertain, 'modest assessments'" that engage citizens as active agents of knowledge, insight, and memory (p. 243; italics in the original).

One concern with the role of non-scientists in the production of science are the implications for established scientific principles.[9] However, as Goodchild (2007) demonstrates, while strictly speaking citizen science might not fulfil scientific criteria per se, it can potentially open up new ways of thinking and approaching data. This is especially relevant for practices of democratisation, which call for different forms of reasoning, as captured in Herbert Simon's (1947) conception of "satisficing" rather than "optimizing" or "maximizing" in decision-making. In opposition to abstractions such as utility theory he advanced an understanding based on how people reason in practice. Practical reasoning, he argued, involves juggling numerous criteria and arriving at a "good enough" solution rather than engaging in an infinite search for all possible ones, evaluating them and then arriving at the best one. Gabrys & Pritchard (2015) take a similar approach to suggest that the adequacy of an answer depends on how practical questions are posed. Instead, they define "just good enough data" to counter the reliance on measurement accuracy as the only objective and criterion for evaluating environmental data gathered through citizen sensing practices. Measurements of environmental phenomena meet different objectives or questions, which are often not known in advance. For instance, a "rough" measurement to identify a pollution event when it is happening or when it has happened might be sufficient and "good-enough". What Gabrys & Pritchard draw attention to is that the potential uses or value of data are often not known in advance and that there is value in organising data production and interpretation as practices of searching for potential rather than reiterating and replicating already known objectives or questions through previously established methods.

Recent experiments by statistical organisations with models of citizen engagement include a pilot project by Statistics Canada using OSM for crowdsourcing citizen work to help fill in data gaps on geolocations (Statistics Canada, 2016).[10] OSM is a collaborative initiative designed to create a free and editable map of the world. The application for Statistics Canada allows users to select a geolocation and edit, for instance, the name of a street. Another example is from the European Commission's Joint Research Centre on Citizen Science and Open Data which has explored possible models of citizen engagement for monitoring the spread of invasive alien plant species (IAS) (Cardoso *et al.*, 2017). That report argued that the implementation of the IAS Regulation could benefit from the contributions of citizens in providing "accurate, detailed, and timely information on IAS occurrences and distribution for efficient prevention, early detection, rapid response, and to allow for evaluation of management measures" (p. 5). Additionally, this form of citizen engagement could raise awareness and increase public support for the regulation as

9. Also see Gabrys et al. *(2016) for discussions about data quality and credibility.*

10. The pilot was organized by Statistics Canada in collaboration with OpenNorth, MapBox, City of Ottawa and OSM Canada. OpenNorth is a non-profit organization developing digital tools for civic engagement.

well as supporting citizens in acquiring skills and better understanding of scientific work (Socientize Consortium, 2014). The United Nations has also identified citizen science data production on environmental issues as necessary to the measurement and monitoring of sustainable development goals (SDGs) (United Nations, 2016). Modes of citizen engagement are recognised as key to ensuring that the 2030 Agenda for Sustainable Development is country-owned and context specific and with goals linked to national values and priorities. While these initiatives conceive of citizen engagement in varying ways, they generally limit it to tasks such as data production, verification and classification. This has led to criticisms of these forms of citizen science as exploitative of citizens as free public labour (DataShift, n.d.; Piovesan, 2017; Paul, 2018). What they point to is that tasks related to data cleaning, coding or analysis as well as design, architecture or interpretation are reserved for experts while citizens are limited to being no more than research subjects or assistants.

We reconceive of citizen science in a way that is more closely aligned with what Jasanoff expresses as the inclusive generation of knowledge. But, following from our argument about detachment, we suggest that inclusivity involves the right to make claims and articulate concerns about how environmental, economic and social issues should be categorised and known.[11] Arguably, this is the claim citizen scientists make when they engage in the independent production of data to challenge or supplement official and scientific knowledge. However, our conception of citizen data envisages citizens not as independent but as co-producers. In this way, we conceive of citizen data as involving new relations between citizens and NSIs in ways that combine statistical science and citizen science. Such a conception could involve citizen engagement in statistical production and lead to statistics that are more representative and inclusive of citizens' concerns, needs and experiences, as well as their own identifications. As such, it would necessitate an approach that is flexible and experimental in its criteria (Paul, 2018) so that it can adapt to the shifting needs and requirements of not only citizens, but also what matters to them. As we suggest below, this includes broadening the understanding of ethics beyond consent, fairness, and data protection to what is arguably at the core of the rise of citizen science: citizens as active in the making and shaping of the data through which official statistics and knowledge are generated. In the next section, we explore what this understanding of ethics might mean in relation to another matter of concern: proposals for "smart statistics".

Smart Statistics

Propositions by Eurostat for the development of "smart statistics" build on conceptions of "smart cities", usually understood as the use of Big Data, urban sensors, Internet of Things (IoT) and other forms of data production and data integration to streamline municipal governance and transportation infrastructures, rejuvenate local economies, transform the urban environment to make it more sustainable, liveable, and socially inclusive (see for instance Henriquez, 2016). While smart cities have been defined in various ways, the concept generally refers to on the one hand how "cities are increasingly composed of and monitored by pervasive and ubiquitous computing and, on the other, whose economy and governance is being driven by innovation, creativity and entrepreneurship, enacted by smart people" (Kitchin, 2014, p. 1). In this view, Big Data offers the possibility of real-time analysis of city life, new modes of urban governance, and envisioning and making more efficient, sustainable, competitive, productive, open and transparent cities.

Leveraging "smart systems" such as smart energy, smart meters, smart transport, and so on is an objective of proposals for "smart statistics" put forward by Eurostat's Big Data Task Force. The proposals seek to engage with the potential of the proliferation of digital devices and sensors connected to the internet and how the data they generate might be embedded in statistical production systems such that statistics could be produced in "real-time" and "automatically".[12] In this view, data capturing, analysis and processing are envisioned as embedded in activities that generate and simultaneously analyse data. The adoption of such an approach could dramatically transform the production system for official statistics and calls for rethinking business processes and architectures, laws and regulations, ethics, methodologies, and so on.

11. This is an understanding advanced in the field of critical citizenship studies and summarised in Isin & Ruppert (2015) and Isin & Saward (2013). Being a citizen is understood as a political subjectivity that includes not only the possession of rights but the right to make rights claims such as the right to shape how data is made about them and the populations of which they are being constituted as a part (Ruppert, 2018).

12. Eurostat Big Data Task Force (2016) "Smart Statistics". Draft document. October.

Two approaches for generating smart statistics understood in this way have been proposed: using third party systems that exist for other purposes than statistics but from which statistical information can be extracted (e.g., mobile phones); or developing entirely new data production practices such as sensors and digital devices exclusively for generating statistical information.[13] The third-party approach engenders many of the concerns we previously identified such as data access and ownership, privacy and ethics, data representativeness, quality, and trust as well as greater detachment between citizens and NSIs. However, the latter approach of designing new devices of data production, provides an opportunity to mitigate these issues. That is, we reconceive of smart statistics as not only requiring that NSIs rethink the technical and organisational aspects of statistical production systems, but also their relations to citizens. As noted in the discussion of citizen science, this could involve models of co-production that engage citizens in the production of smart statistics.

It would, however, mean being care-full in the ways we previously outlined including a broader understanding of ethics that extends throughout the production of official statistics. Ethics of course have long been central principles of official statistics, which address the values of utility, professional standards and ethics, scientific principles, transparency, quality, timeliness, costs, respondent burden, and confidentiality (UN, 2014).[14] These principles constitute what we would call an ethic of care for data, such as care for the quality, accessibility and clarity of data, but also for relations and accountabilities to citizens through practices such as data protection, confidentiality, consent, and trust. While the origins of these principles are a mix of legal, governmental, political and professional rationales and requirements, they tend to operate as part of everyday working values and commitments. This is evident in claims made by statisticians such as "just because you can, doesn't mean you should" use Big Data sources.

The fundamental principles of official statistics thus express a broad conception of ethics that includes relations to citizens that social science research calls procedural ethics (Guillemin & Gillam, 2004). Procedural ethics are understood as an estimation of the ethical issues that might be involved when research and data production are undertaken. However, Guillemin and Gillam note a second dimension of ethics in research, which they term "ethics in practice" (id., p. 261). It concerns the recurrent, iterative, and uncertain ethical moments that happen during research and which may be odds with that covered in a procedural ethics review. This latter understanding is relevant to practices involved in the co-production of smart statistics, which, by definition, involve uncertainty, adaptation and responsiveness to the interactions, interests and demands of different stakeholders. As such, co-production demands an ethic of care that recognises and is responsive to the dependence on relations to citizens and their labours to "create, hold together and sustain" data (Puig de la Bellacasa, 2012, p. 198).

The concept of citizen data we propose thus reconceives of smart statistics as involving new relations to citizens as co-producers of data production platforms. It is a conception that calls for a care-full approach that enlarges the understanding of ethics to include the demands, interests and contributions of citizens at different stages of the development of new devices of data production rather than at the backend as an afterthought or correction. As such, it is a model that builds on the premises of another matter of concern, privacy-by-design, which addresses issues of privacy and consent at the frontend of software design, which we address next.

Privacy-by-Design

Big Data and new data sources come with new questions concerning privacy, consent and confidentiality that are not always fully addressed by existing regulatory frameworks. As such, privacy-by-design has become as matter of concern for NSIs. Privacy-by-design is understood as the embedding of privacy protection at the software design stage of data production platforms, devices or applications. It entails designing privacy protection with citizens in mind at the outset and the implementation of these designs in a

13. Ibid. *One example is Statistics Netherlands collection of data for statistics about road traffic intensities which are produced purely on the basis of road sensors. See: https://www.cbs.nl/en-gb/our-services/innovation/nieuwsberichten/recente-berichten/new-steps-in-big-data-for-traffic-and-transport-statistics.*

14. *Six principles are that: official statistics must meet the test of practical utility; be developed according to strictly professional considerations, scientific principles and professional ethics; present information on the scientific standards of their sources, methods and procedures; may be generated from all types of sources such as surveys or administrative records and the source chosen with regard to quality, timeliness, costs and the burden on respondents; are to be strictly confidential and used exclusively for statistical purposes; and the laws, regulations and measures governing them should be public.*

transparent manner. As such, privacy-by-design is a response to the problem of privacy, consent, and confidentiality through software and which can be used in tandem with other tools, such as privacy impact assessments. By employing privacy-by-design, privacy issues are addressed at the beginning of the design process, in contrast to other approaches that aim at solving privacy issues after software development is complete or leave privacy considerations to legal or regulatory frameworks.

Cavoukian *et al*. (2010) define privacy-by-design through seven foundational principles: proactive not reactive and preventative not reactive; privacy as the default; privacy embedded into design; full functionality that leads to positive sum, not zero-sum outcomes; end-to-end lifecycle protection; visibility and transparency; and respect for user privacy. These principles require designs to be committed to privacy from the beginning and to limit data production to ways that are respectful of citizens' expectations. The principles also require that data production software addresses the likelihood that data may exist after the software stops functioning. The authors also emphasise that the lifecycle of software must be considered when deciding on how to best protect privacy, including making plans for deleting data once the software reaches the end of its lifecycle. Finally, the principles compel organisations dealing with personal data to be transparent in their goals and to remain accountable to citizens.

However, the production and processing of personal data present many other challenges for privacy in addition to individual privacy. Nissenbaum (2004) argues that privacy norms need to be tied to specific contexts. She describes three principles that have dominated debates around privacy throughout the 20th century, namely, limiting surveillance of citizens by governments, restricting access to private information, and curtailing intrusions into private places. She suggests a new term, "contextual integrity", to deal with the new challenges introduced by digital technologies. Contextual integrity demands that information gathering is kept appropriate to the context and obeys the governing norms of distribution within it. The key insight is that norms of distribution vary across cultures, historical periods, locales, and other factors. Additionally, contextual integrity requires awareness of not only the specific site of data production but also the relevance of related social institutions (Nissenbaum, 2009).

Approaches that aim to protect individual privacy may still lead to undesired outcomes in large-scale data production efforts. When individually anonymised data are joined to create profiles, individuals who fit the profile could still experience effects even when they are not identified individually. For example, Graham (2005) discusses how software can be used to assign different categories to different parts of a city based on school performance, house prices, crime rates, etc., which might potentially orchestrate inequalities and discriminate inhabitants, even when they are not personally identified. Similarly, Zwitter (2014) has identified and problematised the potential discriminatory "group effects" of anonymised data such as in practices of profiling.

The use of Big Data also introduces additional privacy challenges. Barocas and Nissenbaum (2014) argue that anonymity and consent are often fundamentally undermined in Big Data applications, and that other approaches are needed to protect integrity, such as policies based on moral and political principles that serve specific contextual goals and values. Instead of focusing on anonymity in Big Data applications, they instead emphasise securing informed consent, not only as a choice for subjects to waive consent or not, but a requirement that data collectors justify their actions in relation to norms, standards, and expectations. To an extent this is addressed in the recently implemented General Data Protection Regulation (GDPR) in member states across the European Union, which is based on a broad understanding of personal data and privacy and will end practices of general consent by default for the production of personal data.[15] It introduces the requirement to think "what is personal data" for all private and public stakeholders which demand, hold or archive personal data, as well as what are the ethical practices required to deal with personal data, given the complexity and connectedness of data systems and proven non-neutrality of algorithms. In sum, privacy is not a single thing but depends on the context of production, accountability for group effects, and mechanisms of informed consent.

Recently, scholars working to address the technical challenges of privacy in relation to Big Data have proposed a method of privacy protection by taking advantage of blockchain technology (Montjoye *et al*., 2014; Zyskind

15. The General Data Protection Regulation came into force in May 2018. See: https://www.eugdpr.org/.

et al., 2015). Blockchain is a distributed computing method where many devices communicate with one another over a shared network, without requiring a central server to authorise the participation of each member or to keep a list of currently connected members. By applying blockchain technology to privacy, it becomes possible to encrypt and distribute private data over a large network without requiring a trusted central server.

Blockchain privacy methods are intended to solve underlying privacy challenges using a technical framework during software development. However, as we have indicated above, they do not stand on their own as the sole solution to ensuring privacy, but rather supplement legal and policy-oriented considerations such as contextual integrity, group effects and modes of consent through software design. We thus reconceive of privacy-by-design beyond software to include citizen privacy as a right that should be built into not only the frontend of software design but through relations with citizens as co-producers in the production of official statistics. That is, like ethics, privacy is processual and cannot be settled through the one-time granting of consent or software design alone or independent of specific contexts.

* *
*

In sum, we have taken up matters of concern expressed by statisticians and reconceived of them as principles of citizen data. Through the discussion of the four principles of experimentalism, citizen science, smart statistics and privacy-by-design, we have explored how citizen data can create new attachments and relations between citizens and NSIs, and between citizens' actions, identifications and experiences and how they are categorised, included and excluded in statistics. In this regard, we argue it has the potential to produce new statistical variables desired and identified by citizens, increase their identification with official statistics and possibly advance their role as also users of statistics. Indeed the latter may well be a collateral effect of co-producing statistics with citizens in ways that are more in accordance with their experiences and knowledge.

We place the significance of our concept of citizen data within the current proliferation of data production platforms that enable myriad data generators (e.g., platform owners) and analysts (e.g., researchers, governments, media) to produce statistics and knowledge of societies (Ruppert *et al*., 2013). Indeed, many topics of interest to NSIs such as price levels, the economy, consumer sentiment or tourism can be measured using Big Data generated by browsers, social media or devices such as mobile phones that can be accessed and analysed by different actors. Some would claim that this represents a "democratisation" of knowledge and the erosion of validated knowledge and expertise about societies. However, as Ruppert *et al*. (2013) contend, this widening distribution of data and analysis means that knowledge of societies does not cohere in single authoritative accounts to the same extent that it perhaps did in the recent past. Instead, what constitutes legitimate knowledge and expertise have become major sites of political contention and struggle as revealed in current debates about "alternative facts".

Proposals that NSIs need to thus defend the quality and legitimacy of official statistics through gatekeeping practices such as demonstrating their trustworthiness by making their statistical practices transparent and thus assessable, fact checking competing statistics, and "calling out bad numbers" certainly have a role to play. However, they potentially play into the premise that what is at stake is winning a competition of "facts". They ignore that what constitutes "public facts" should be open to democratic contestation and deliberation because they inevitably involve normative judgements about social meaning and choices about which experiential realities matter (Jasanoff & Simmit, 2017). We thus suggest NSIs have a role to play in fostering official statistics as social and collective accomplishments where their legitimacy is derived from conditions of co-production that address data subjects as citizens with rights to be active participants. Such an approach understands data and official statistics as social technologies that require new forms of engagement and relations between experts, decision-makers, and citizens for addressing collective problems (Jasanoff, 2003) and as matters of democratic deliberation where citizens are active in the making and shaping of knowledge about societies of which they are a part.

We recognise that the concept of citizen data raises many practical and political questions. For one, we are not suggesting that existing methods and their relations to citizens will

become obsolete. However, methods such as surveys and questionnaires will likely change as digital technologies are increasingly adopted and a concept of citizen data can possibly inform those changes. That is, beyond Big Data sources, how data is produced by NSIs using various methods can be reconceived along the lines of what we call citizen data. While online or digital surveys and censuses, for example, are being adopted they do not imagine the possibilities of co-production. Different modes of co-production could be adopted that utilise the affordances of digital technologies and potentially produce data that more closely aligns with the experiences and knowledge of citizens.

Throughout our discussion we have defined co-production as involving citizens in the statistical production process. What this would mean practically is of course a major question and extends to issues of representativeness and inclusion in that process. This is a matter of concern for all methods especially taking into account the heterogeneity of citizens. For methods that mobilise digital technologies such as online censuses and surveys this is potentially exacerbated by what has come to be called the "digital divide". These are only some of the possible practical and political issues that arise from citizen data, which we also addressed in the collaborative workshop with statisticians noted previously. While we have not reported on the outcomes of that workshop in this article, one outcome was imagining alternative "roadmaps" for engaging with citizens at different stages of the statistical production process, from the co-design of prototypes for data generation platforms and apps to the establishment of co-operative forms of data ownership. In other words, citizen data does call for rethinking statistical production processes and some of their fundamental premises.

For example, aspects of statistical production that would need to be rethought are those of data standards and quality. However, as noted, the principle of experimentalism calls for being open to such questions and not settling them in advance including what may or could constitute quality. Interestingly, this is also recognised in NSI experiments with Big Data generated by third party systems where concerns about quality as well as others such as the representativeness of data have been raised. One solution statisticians propose is that statistics that repurpose Big Data could be adopted not as replacements but auxiliary, complementary or supplementary to existing data sources. While possibly relegating such data to a different status and role, this response provides an opportunity to rethink how statistics are made "official". That is, it suggests that there is not one mode of production or set of standards through which data can be made official. We suggest that this also applies to existing methods that produce data for official statistics but which involve myriad standards and where quality is not singularly defined or measurable. However, the concept of citizen data that we have developed introduces a critical difference that goes beyond issues of standards and quality. It proposes that the authority and expertise to make statistics official are not founded in a single institution, but in processes of co-production and direct relations to citizens. In that regard, citizen data approaches claims of "alternative facts" as not matters of accuracy and standards but of the relations to citizens through which data and in turn statistics are made official. □

BIBLIOGRAPHY

Barocas, S. & Nissenbaum, H. (2014). Big Data's End Run Around Anonymity and Consent. In Lane, J., Stodden, V. Bender, S. & Nissenbaum, H. (Eds.), *Privacy, Big Data, and the Public Good*, pp. 44–75. Cambridge, MA: Cambridge University Press.

Binder, T., Brandt, E., Ehn, P. & Halse, J. (2015). Democratic Design Experiments: Between Parliament and Laboratory. *CoDesign*, 11(3-4), 152–165. https://doi.org/10.1080/15710882.2015.1081248

Boltanski, L. & Chiapello, E. (2007). *The New Spirit of Capitalism*. London: Verso.

Cardoso, A. C., Tsiamis, K., Gervasini, E. *et al.* (2017). Citizen Science and Open Data: a model for Invasive Alien Species in Europe. Joint Research Centre (JRC) and the European Cooperation in Science and Technology (COST Association), *Workshop Report*. Brussels, BE. https://doi.org/10.3897/rio.3.e14811

Callon, M., Burchell, G., Lascoumes, P. & Barthe, Y. (2011). *Acting in an Uncertain World: An Essay on Technical Democracy*. Cambridge, MA: MIT Press.

Cavoukian, A., Taylor, S. & Abrams, M. E. (2010). *Privacy by Design*: Essential for Organizational Accountability and Strong Business Practices. *Identity in the Information Society*, 3(2), 405–413. https://doi.org/10.1007/s12394-010-0053-z

DataShift (n.d.). Global Goals for Local Impact: Using Citizen-Generated Data to Help Achieve Gender Equality. http://civicus.org/thedatashift/wp-content/uploads/2017/01/LanetUmojaProcessandApproach.pdf (accessed 22 February 2018)

European Commission (2013). Environmental Citizen Science. *Science for Environment Policy Indepth Report* N° 9. Bristol: University of the West of England, Science Communication Unit. http://ec.europa.eu/environment/integration/research/newsalert/pdf/IR9_en.pdf (accessed 22 February 2018)

Fleurbaey, M. (2009). Beyond GDP: The Quest for a Measure of Social Welfare. *Journal of Economic literature*, 47(4), 1029–1075. https://doi.org/10.1257/jel.47.4.1029

Gabrys, J., Pritchard, H. & Barratt, B. (2016). Just Good Enough Data: Figuring Data Citizenships Through Air Pollution Sensing and Data Stories. *Big Data & Society*, 3(2), 1–14. https://doi.org/10.1177/2053951716679677

Gabrys, J. & Pritchard, H. (2015). Just Good Enough Data and Environmental Sensing: Moving Beyond Regulatory Benchmarks toward Citizen Action. In *Infrastructures and Platforms for Environmental Crowd Sensing and Big Data*. Barcelona: European Citizen Science Association. https://ecsa.citizen-science.net/sites/default/files/envip-2015-draft-binder.pdf (accessed 22 February 2018)

Goodchild, M. F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Graham, S. (2005). Software-Sorted Geographies. *Progress in Human Geography*, 29(5), 562–580. https://doi.org/10.1191/0309132505ph568oa

Grommé, F., Ustek-Spilda, F., Ruppert, E. & Cakici, B. (2017). Citizen Data and Official Statistics: Background Document to a Collaborative Workshop. ARITHMUS *Working Paper* N° 2. http://arithmus.eu/wp-content/uploads/2015/02/ARITHMUS-collaborative-workshop-wp_final-version-060717-1.pdf

Grommé, F. (2015). *Governance by Pilot Projects: Experimenting with Surveillance in Dutch Crime Control* (Doctoral thesis). Amsterdam: University of Amsterdam. http://hdl.handle.net/11245/1.486712 (accessed 6 February 2019)

Guillemin, M. & Gillam, L. (2004). Ethics, Reflexivity, and "Ethically Important Moments" in Research. *Qualitative Inquiry*, 10(2), 261–280. https://doi.org/10.1177/1077800403262360

Henriquez, L. (2016). *Amsterdam Smart Citizens Lab: Towards Community Driven Data Collection*. Amsterdam: De Waag Society and AMS Institute. https://waag.org/sites/waag/files/media/publicaties/amsterdam-smart-citizen-lab-publicatie.pdf (accessed 2 April 2017)

Hussmanns, R. (2004). Measuring the Informal Economy: From Employment in the Informal Sector to Informal Employment. *Working Paper* N° 53. http://www.ilo.org/wcmsp5/groups/public/---dgreports/---integration/documents/publication/wcms_079142.pdf (accessed 30 April 2018)

Isin, E. & Ruppert, E. (2015). *Being Digital Citizens*. London: Rowman & Littlefield International.

Isin, E. & Saward, M. (2013). *Enacting European Citizenship*. Cambridge: Cambridge University Press.

Jasanoff, S. (2003). Technologies of Humility: Citizen Participation in Governing Science. *Minerva*, 41(3), 223–244. https://doi.org/10.1023/A:1025557512320

Jasanoff, S. & Simmet, H. R. (2017). No Funeral Bells: Public Reason in a "post-Truth" Age. *Social Studies of Science*, 47(5), 751–770. https://doi.org/10.1177/0306312717731936

Jungnickel, K. (2017). Making Things to Make Sense of Things: DIY as Research Subject and Practice. In: Sayers, J. (Ed.), *The Routledge Companion to Media Studies and Digital Humanities*, pp. 492–502. Oxon: Routledge.

Kitchin, R. (2014). The Real-Time City? Big Data and Smart Urbanism. *GeoJournal*, 79(1), 1–14. https://doi.org/10.1007/s10708-013-9516-8

Kullenberg, C. & Kasperowski, D. (2016). What Is Citizen Science? – A Scientometric Meta-Analysis. *PLOS ONE*, 11(1), e0147152. https://doi.org/10.1371/journal.pone.0147152 (accessed 2 April 2017)

Latour, B. (2004). Why Has Critique Run Out of Steam? From Matters of Fact to Matters of Concern. *Critical Inquiry*, 30(2), 225–248. https://doi.org/10.1086/421123

Latour, B. (2006). Which Protocol for the New Collective Experiments? *Boletín CF+S*, (32/33). http://habitat.aq.upm.es/boletin/n32/ablat.en.html (accessed 2 April 2017)

Marres, N. (2012). *Material Participation: Technology, the Environment and Everyday Publics*. Basingstoke: Palgrave Macmillan.

Montjoye, Y.-A. (de), Shmueli, E., Wang, S. S. & Pentland, A. S. (2014). OpenPDS: Protecting the Privacy of Metadata through SafeAnswers. *PLOS ONE*, 9(7). https://doi.org/10.1371/journal.pone.0098790

Muniesa, F. & Linhardt, D. (2011). Trials of Explicitness in the Implementation of Public Management reform. *Critical Perspectives on Accounting*, 22(6), 550–566. https://doi.org/10.1016/j.cpa.2011.06.003

Nissenbaum, H. (2004). Privacy as Contextual Integrity. *Washington Law Review*, 79(1), 119–158. https://nyuscholars.nyu.edu/en/publications/privacy-as-contextual-integrity (accessed 2 April 2017)

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford: Stanford University Press.

Paul, K. T. (2018). Collective Organization of Discourse Expertise Using Information Technology – CODE IT! *Information Technology*, 60(1), 21–27. https://doi.org/10.1515/itit-2017-0022

Piovesan, F. (2017). *Statistical Perspectives on Citizen-Generated Data*. [Online]. http://civicus.org/thedatashift/wp-content/uploads/2015/07/statistical-perspectives-on-cgd_web_single-page.pdf (accessed 22 February 2018)

Puig de la Bellacasa, M. (2012). "Nothing Comes Without Its World": Thinking with Care. *The Sociological Review*, 60(2), 197–216. https://doi.org/10.1111/j.1467-954X.2012.02070.x

Rabinow, P. & Bennett, G. (2012). *Designing Human Practices: An Experiment with Synthetic Biology*. Chicago: University of Chicago Press.

Ruppert, E. (2018). *Sociotechnical Imaginaries of Different Data Futures: An Experiment in Citizen Data*. 3e Van Doornlezing. Rotterdam, NL: Erasmus School of Behavioural and Social Sciences. https://www.eur.nl/sites/corporate/files/2018-06/3e%20van%20doornlezing%20evelyn%20ruppert.pdf (accessed 21 Jan 2019)

Ruppert, E., Law, J. & Savage, M. (2013). Reassembling Social Science Methods: The Challenge of Digital Devices. *Theory, Culture & Society, Special Issue on "The Social Life of Methods"*, *30*(4), 22–46. https://doi.org/10.1177/0263276413484941

Ruppert, E., Harvey, P., Lury, C., Mackenzie, A., McNally, R., Baker, S. A., Kallianos, Y. & Lewis, C. (2015). A Social Framework for Big Data. Project Report. CRESC, The University of Manchester and The Open University. http://research.gold.ac.uk/13483/ (accessed 2 April 2017)

Simon, H. (1947). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. New York: Macmillan.

Socientize Consortium (2014). Green Paper on Citizen Science. Citizen Science for Europe: Towards a Better Society of Empowered Citizens and Enhanced Research. European Commission Digital Science Unit. https://ec.europa.eu/digital-single-market/en/news/green-paper-citizen-science-europe-towards-society-empowered-citizens-and-enhanced-research (accessed 22 February 2018)

Statistics Canada (2016). *Open Building Data: An Exploratory Initiative*. http://www.statcan.gc.ca/eng/crowdsourcing (accessed 18 February 2018)

Stengers, I. (2010). *Cosmopolitics*. Vol. 1–2. Minneapolis: University of Minnesota Press.

Stiglitz, J. E., Sen, A. & Fitoussi, J.-P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*. Paris: CMESP. http://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+Commission+report (accessed 30 April 2018)

Struijs, P., Braaksma, B. & Daas, P. J. H. (2014). Official statistics and Big Data. *Big Data & Society*, 1(1), 1–6. https://doi.org/10.1177/2053951714538417

UNECE (2014). The Role of Big Data in the Modernisation of Statistical Production Project. Report of the Big Data Privacy Task Team. http://bit.ly/2eTHDOe (accessed 2 April 2017)

United Nations (2014). "Fundamental Principles of Official Statistics". *Resolution adopted by the General Assembly on 29 January 2014. A /RES/68/261*. http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf (accessed 21 Jan 2019)

United Nations (2016). Make Sustainable Development Goals Relevant to Citizens. New York: Economic and Social Council. https://www.un.org/press/en/2016/ecosoc6782.doc.htm (accessed 2 April 2017)

Waterton, C. & Tsouvalis, J. (2015). On the Political Nature of Cyanobacteria: Intra-Active Collective Politics in Loweswater, the English Lake District. *Environment and Planning D: Society and Space*, 33(3), 477–493. https://doi.org/10.1177/0263775815594305

Zwitter, A. (2014). Big Data Ethics. *Big Data & Society*. 1(2) 1–6. https://doi.org/10.1177/2053951714559253

Zyskind, G., Nathan, O., & Pentland, A. (2015). Decentralizing Privacy: Using Blockchain to Protect Personal Data. *2015 IEEE Security and Privacy Workshops,* pp. 180–184. https://doi.org/10.1109/SPW.2015.27