An ethico-legal framework for social data science

Forgó, Nikolaus; Hänold, Stefanie; van den Hoven, Jeroen; Krügel, Tina; Lishchuk, Iryna; Mahieu, René; Monreale, Anna; Pedreschi, Dino; Pratesi, Francesca; van Putten, David

DOI
10.1007/s41060-020-00211-7

Publication date
2020

Document Version
Final published version

Published in
International Journal of Data Science and Analytics

Citation (APA)
Forgó, N., Hänold, S., van den Hoven, J., Krügel, T., Lishchuk, I., Mahieu, R., Monreale, A., Pedreschi, D., Pratesi, F., & van Putten, D. (2020). An ethico-legal framework for social data science. *International Journal of Data Science and Analytics*, *11*(4), 377-390. https://doi.org/10.1007/s41060-020-00211-7

International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-020-00211-7

REGULAR PAPER

An ethico-legal framework for social data science

Nikolaus Forgó[1] · Stefanie Hänold[2] · Jeroen van den Hoven[3] · Tina Krügel[2] · Iryna Lishchuk[2] · René Mahieu[3,4] · Anna Monreale[5] · Dino Pedreschi[5] · Francesca Pratesi[5,6] · David van Putten[3]

Received: 1 August 2019 / Accepted: 10 March 2020

Abstract
This paper presents a framework for research infrastructures enabling ethically sensitive and legally compliant data science in Europe. Our goal is to describe how to design and implement an open platform for big data social science, including, in particular, personal data. To this end, we discuss a number of infrastructural, organizational and methodological principles to be developed for a concrete implementation. These include not only systematically tools and methodologies that effectively enable both the empirical evaluation of the privacy risk and data transformations by using privacy-preserving approaches, but also the development of training materials (a massive open online course) and organizational instruments based on legal and ethical principles. This paper provides, by way of example, the implementation that was adopted within the context of the SoBigData Research Infrastructure.

Keywords Ethical data science · Legal data science · Research infrastructure

1 Introduction

In the last years, we have witnessed different initiatives in Europe aimed at providing environments and infrastructures to share research data and technologies, in accordance with the principles of Open Research Data and Open Science. The general idea of these initiatives is to provide ecosystems for enhancing scientific collaborations among researchers and practitioners, even those from different disciplines. Examples of recent initiatives in different research fields are: EOSCpilot and SoBigData (social sciences), SeaDataCloud (environmental and earth sciences), IN-SKA (physical sciences), EVAg (biological and medical sciences).

In the field of ICT, the SoBigData Research Infrastructure (RI) aims at providing a platform or ecosystem for ethics-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, or, in other words, social big data science. More and more often, these data regard private aspects of our lives, such as

This work was supported by the European Commission through the Horizon2020 European project "SoBigData Research Infrastructure—Big Data and Social Mining Ecosystem" (Grant Agreement 654024). The funders had no role in developing the research and writing the manuscript.

✉ Francesca Pratesi
francesca.pratesi@isti.cnr.it

Nikolaus Forgó
nikolaus.forgo@univie.ac.at

Stefanie Hänold
stefanie.haenold@iri.uni-hannover.de

Jeroen van den Hoven
M.J.vandenHoven@tudelft.nl

Tina Krügel
kruegel@iri.uni-hannover.de

Iryna Lishchuk
iryna.lishchuk@iri.uni-hannover.de

René Mahieu
Rene.Mahieu@vub.be

Anna Monreale
anna.monreale@unipi.it

Dino Pedreschi
pedre@di.unipi.it

David van Putten
davidvanputten@gmail.com

[1] Universität Wien, Vienna, Austria

[2] Leibniz Universität Hannover, Hannover, Germany

[3] Delft University of Technology, Delft, The Netherlands

[4] Vrije Universiteit Brussel, Brussels, Belgium

[5] University of Pisa, Pisa, Italy

[6] ISTI-CNR, Pisa, Italy

Published online: 31 March 2020

our movements [15], healthcare [28], our social interactions and emotions [13].

SoBigData, as most of the RI, follows the cloud paradigm that enables users and organizations to have access to a virtually unlimited amount of resources to store, manage and process data in a reliable infrastructure. The cloud paradigm has different advantages ranging from scalability to easy access to data and applications, from cost savings on equipment to easy collaborations among different locations. This model certainly revolutionized the way to consume IT in the world, and the benefits of this model have been also recognized at the European level. Cloud environments are clearly considered a suitable means to share research data and thus for enhancing scientific research. In this particular case, the cloud infrastructure might enable the fast sharing of (personal) data, which allows for collaboration among researchers from different disciplines.

SoBigData also strives to formulate a vision of responsible data science for the social sciences and associated innovations that is fit for the twenty-first century. Since SoBigData is focused on topics of social data mining, it typically concerns the processing of personal data, including *sensitive* personal data, describing human individuals and activities. In this context, it therefore becomes fundamental to take into consideration the legal and ethical aspects of processing of personal data, especially given the entry into force of the General Data Protection Regulation (GDPR) in May 2018. This is why it aims at devising methodologies and tools that enable us to arrive at data science solutions that are demonstrably in accordance with shared societal and moral values. For this reason, it is important that legal requirements and constraints are complemented by a solid understanding of ethical and legal views and values such as privacy and data protection.

This approach exemplifies the basic idea of Responsible Research and Innovation (RRI), which is given a prominent role by the European Commission in the Horizon 2020 Program. This idea was embraced by the EU, first partially in the Lund declaration (2009) [9] and then comprehensively in the Rome declaration on RRI (2014) [27]. Central to this idea is that innovations in all fields should aim at solving societal or global problems, without creating bigger problems than the ones they are trying to solve and also, ideally, accommodating a number of moral values that are in relations of mutual tension, e.g. safety and efficiency, open research data and privacy, sustainability and prosperity.

Aiming at solving societal problems is a core condition for the legitimacy of *big data science*, and demonstrable compliance with core moral values is crucial for the warranted trust that citizens should be able to place in data science and the applications of big data research. If we succeed in overcoming conflicts between values, by applying value-sensitive design (VSD), we may say that we have successfully harnessed technology and applied science to contribute to moral progress.

The main task of RRI and VSD in the field of big data is to help develop the science and the tools that on the one hand allow users to make use of the functionalities and capabilities that big data can offer to help us solve our problems, while at the same time allowing them to respect fundamental rights and accommodate shared values, such as privacy, security, safety, fairness, equality, human dignity and autonomy. It may be difficult to reconcile each of our moral values with any of the others pairwise at any given time, since there are bound to be tensions and conflicts between them. It is even harder to satisfy all of these moral values and societal demands in one fell swoop. Moreover, we should not give undue emphasis to one value over all the others. This fundamental difficulty should not be ignored, and it should not give rise to complacency or despair. The moral difficulty of reconciling a variety of conflicting moral requirements also constitutes an opportunity for scientific discovery and societal innovations. The very idea of privacy-enhancing technology (PET) is a case in point.

Privacy and human rights are often reduced to obstacles on the road to economic growth and societal success. Zuboff, in her analysis of surveillance capitalism [33], argues that big corporations operate under what she calls a "Street View model"; the maximum possible amount of data is gathered and restriction, if it ever happens, only happens ex post, under pressure of the law. As Chris Anderson, Silicon Valley Tech Entrepreneur, once put it, expressing an attitude common to various organizations: "it is better to ask for forgiveness than permission" [1]. Similarly, what governments all over the world are trying to accomplish in the field of national security and intelligence agencies in terms of wiretapping and profiling has received ample attention in the media. Academia and science now need to lead the way and show that a responsible development and use of big data, namely one that both achieves desired functionality, efficiency and is ethically grounded in shared values and human rights at the same time, is in fact possible.

We hold that the use and utilization of big data can only count as *responsible* if it deals adequately with moral values such as privacy, confidentiality, accuracy, transparency, autonomy, fairness and equal access. In a field that is still far from settled, such as that of ethical use of big data, adequately means at least pushing the state of the art forward. We assume that a scientific discipline, a field of technology or an infrastructure in this field (of which SoBigData is a prime example), cannot be called "responsible" without a coherent account of a methodology or methodological framework, criteria, mechanisms and procedures for (1) systematically assessing the ethical and legal relevance of requirements, assumptions and ramifications at all levels and for all of its applications, (2) diligently designing for moral and value

requirements, (3) promoting data use and data management (via e.g. suitable reputation systems) where accountability can be assigned.

To this end, we have implemented an ethical and legal framework in accordance with European legislations, including data protection and intellectual property right. Moreover, it monitors the compliance of analytical processes and research activities with this framework and provides tools and methodologies for developing of big data analytics and social mining tools with value-sensitive design [31] and privacy-by-design methodologies [21].

Contribution and organization This paper provides an overview of the ethical and legal framework that has been developed within the legal and ethical work package of the European SoBigData project. The presented framework is the result of a joint work between philosophers, lawyers and technical privacy experts. To begin, we will outline the underlying legal framework stipulated by the General Data Protection Regulation and the basic rules concerning intellectual property law (Sect. 2). In the process, we will clarify the legal and ethical responsibilities of the actors within the SoBigData RI. We will then describe how these abstract rules have been translated into a number of concrete practices, which ensure the compliance with legal norms and ethical requirements (Sect. 3). In particular, we will also describe a practical example that shows the application of a SoBigData tool for privacy risk assessment. Lastly, Sect. 4 concludes the paper highlighting and discussing some important open issues.

2 The legal and ethical framework

The processing of various kinds of social data, which include GPS information, social media data and socio-economic information, involves a number of responsibilities and often requires to take care of legal requirements. In the following, we present an overview of the most important aspects of the European data protection framework (Sect. 2.1), of the applicable rules on intellectual property (Sect. 2.2) as well as of the moral and ethical issues raised by the big data age (Sect. 2.3).

2.1 European data protection law

2.1.1 The General Data Protection Regulation

The General Data Protection Regulation (GDPR)[1] is applicable since 25th of May 2018, and it is the most important source of law in Europe that needs to be considered when personal data are processed for research. The overall aim of the Regulation is to protect individuals in their fundamental rights and freedoms and in particular to their right to the protection of their personal data.[2] Hitherto, the processing of personal data has been regulated on the European level by the Data Protection Directive (DPD)[3] which has been enacted in the year 1995 and was transposed into national law throughout the EU.

Such a reform was necessary: since the enactment of the DPD, rapid technological developments and the ongoing globalization have brought new challenges to the protection of personal data. It is not only that the scale of the collection of personal data has risen dramatically, but also the exchange of personal data between public and private actors has massively increased. The new Regulation shall respond to these technical advances and practices and as well align the level of data protection in all the Member States.[4] The General Data Protection Regulation, while upholding old established data protection principles, also promotes new features and gives emerging rules and principles a rigid legal framework. Mayer-Schönberger and Padova [20] described the new Regulation fittingly as an "unusual hybrid of old and new". Due to its nature as a regulation, the rules laid down in the GDPR are directly applicable in all European Member States. However, especially in the research field, the Regulation also provides for a number of implementing acts by the Member States which researchers may also have to consider in the event of processing personal data for their research.[5]

2.1.2 Material and territorial scope of the GDPR

The GDPR framework applies when personal data are processed wholly or partly by automated means.[6] The data protection rules do not apply to the processing of anonymous data.[7] According to Art. 4 No. 1 GDPR, personal data are "any information relating to an identified or identifiable natural person". Thereby an "identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier" such as the name or an identification number.[8] "To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, (…) either by the controller

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[2] Art. 1(2) GDPR.

[3] Directive 95/46/EC of the European Parliament and the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

[4] Recitals 1–13 GDPR.

[5] For example, Art. 89(2) and (3) GDPR and Art. 9(2)(j) GDPR.

[6] Art. 2(1) GDPR.

[7] Recital 26 GDPR.

[8] Art. 4(1) GDPR.

or by another person to identify the natural person directly or indirectly", "such as the costs of and the amount of time required for identification" taking into account the available technology.[9] The apparently convenient way to escape from the compliance with data protection rules by working with anonymized data sets has pitfalls. In the era of big data, it has become more and more difficult to de-personalize data sets in order for them to be considered as anonymous data in a legal sense [2,3,16,18,23,29,30]. In addition, a regular check of the status of the data is required because re-identification procedures that are exorbitantly expensive today may be affordable in the near future due to fast technological developments [17,18,30].

Further regulations regarding the material and territorial scope of the Regulation are to be found in Arts. 2 and 3 GDPR. The European legislator had expanded the territorial scope of application. The Regulation generally applies to the processing of personal data in the context of activities of an establishment of a controller or a processor in the European Union, and it does not matter where the processing takes place, whether the data belong to EU citizens or where the data were collected (Art. 3(1) GDPR). Controllers or processors that are established outside of the EU and who process personal data of data subjects who are in the Union also must comply with the Regulation under certain circumstances, e.g. when they monitor behaviour of people as far as it takes place in the EU.[10] At this development stage, the SoBigData RI is limited to data processing in an European context. Transeuropean transfers of personal data are of particular complexity as not only Arts. 44–50 GDPR become relevant, but also regulations of third countries will have to be taken into account. In the future SoBigData will strive for enabling international transfers of data, but this requires to adapt the framework to the respective legal requirements.

2.1.3 Data protection principles

In Art. 5, the Regulation stipulates the principles relating to processing of personal data, which the data controller shall be responsible for and also be able to demonstrate compliance with (accountability[11]). These principles require, for example, that personal data shall be processed lawfully, fairly and in a transparent manner (principle of lawfulness, fairness and transparency). Additionally, personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes (principle of purpose limitation). Furthermore, only personal data that are adequate, relevant and limited to what is necessary for the processing purpose shall be processed (principle of data minimization). Moreover, personal data must be accurate and where necessary kept up to date (principle of accuracy). Personal data shall as well only be stored as long as the processing purpose requires it (principle of storage limitation). It is also required that the integrity and confidentiality of the personal data is appropriately secured (principle of integrity and confidentiality). The Regulation contains a number of articles that further specify these principles, e.g. Art. 6 and 9 GDPR entail provisions that determine under which circumstances the processing of personal data is lawful. It must be verified for each individual research project whether a legal ground applies or not. One legal ground may be the (explicit) informed consent given by the data subject [Art. 6 (1)(a) or Art. 9 (2)(a) GDPR]. However, often there will be no informed consent. In these cases, it must be evaluated whether Art. 6(1)(f) or Art. 9(2)(j) GDPR in conjunction with the national research exemptions constitutes a legal ground for the envisaged processing of personal data.

2.1.4 Privacy by design and by default

Another manifestation of these general data protection principles is the tenet of data protection by design and by default, contained in Art. 25 GDPR. This provision is one of the core elements and one of the innovation drivers of the Regulation [19]. Its underlying concepts of privacy by design and by default, however, have been propagated in practice and the scientific literature for quite some time now [14,19]. The core principle of data protection and privacy by design is that privacy and data protection should be built into the design and architecture of technologies. This should happen with respect to technical as well as organizational aspects [8]. Privacy-friendly engineering and organizational arrangements are estimated as a prerequisite for creating more transparency for data subjects regarding the collection and further processing of their information. Transparency, in turn, is a fundamental requirement for data subjects to exercise control over their data [8] and also for increasing trust on their side.

According to Art. 25 of the Regulation, the data controller is legally obliged to implement technical and organizational measures in order to meet the legal requirements of the Regulation, especially the above-mentioned data protection principles entailed in Art. 5 GDPR, and generally to protect the rights of data subjects to data protection. This tenet applies as soon as the purpose of the processing of the personal data is set, but must be continuously followed during all processing steps. Type and scale of the required action to be taken by the data controller depend on the state of the art, the cost of implementation and the nature, scope and context and purposes of the processing as well as the risks posed by the processing for the concerned individuals. According to the principle of data protection by default, the data controller

[9] Recital 26 GDPR.

[10] Art. 3(2)(b) GDPR.

[11] The tenet of accountability is explicitly mentioned in Art. 5(2) GDPR.

shall implement measures for ensuring that by default only personal data that are necessary for each specific purpose are processed. In the context of research, Art. 89 (1) GDPR should be also taken into consideration, which describes specific rules with regard to the principle of data minimization.

2.1.5 The rights of data subjects

The General Data Protection Regulation strongly emphasizes the rights of data subjects, enabling individuals to enforce their right to protection of their personal data. The rights are to be found in chapter III of the Regulation and include comprehensive rights to information (Arts. 13 and 14 GDPR) and access (Art. 15 GDPR). According to Arts. 13 and 14 GDPR, the data controller has to inform the data subject, inter alia, about the purpose of the processing and the categories of personal data concerned. The controller must, e.g. also provide information about his or her identity and the legal basis for the processing. The information obligations are in total fairly comprehensive and aim to give the data subject an understanding about the processing of their personal data.

The Regulation acknowledges that there are circumstances where these information requirements put too much of a burden on the data controller. For example, where personal data have not been obtained directly from the data subject, Art. 14(5)(b) GDPR provides an exception if the provision of such information proves impossible or would involve a disproportionate effort or if the information obligation is likely to render impossible or seriously impair the aim of the processing.[12] In these cases, though, the Regulation requires that the data controller takes appropriate measures to protect the data subject's rights, freedoms and legitimate interests. One measure explicitly mentioned by the law is "making the information publicly available". This does not ensure that the concerned data subject becomes aware of the data processing, but public attention and awareness by the data protection authorities are made possible. Art. 11 GDPR also has a restrictive effect on the obligations on side of the data controllers in cases where controllers are not in the position to identify the data subject with the data at their disposal. Data controllers are accordingly not obliged to maintain, acquire or process additional information in order to identify the data subject only for complying with their duties under the Regulation, e.g. the information obligations. Another right of the data subject provided is the "right to be forgotten" which gives the data subject under certain circumstances the right to erasure of personal data concerning him or her (Art. 17(1) GDPR), e.g. when consent is withdrawn. If the data processing is based on the informed consent of the data subject, this consent can be withdrawn by the data subject at any time (Art. 7(3) GDPR). However, research is also privileged in this context, as Art. 17(3)(d) GDPR provides restrictions to the "right to be forgotten" if its exertion is likely to render impossible or seriously impair the achievement of the research purpose.

2.1.6 Who is responsible for complying with data protection law?

It is the data controller who is mainly responsible for taking care that the processing of personal data is in compliance with the data protection provisions, e.g. the general principles contained in Art. 5(1) GDPR [see 3 (a.3)]. The Regulation provides explicitly in Art. 5(2) that the controller shall be responsible for and be able to demonstrate compliance with these principles. Art. 25 GDPR, stipulating the tenet of privacy by design and by default, also applies to the data controller. The data subject's rights, e.g. the right to information[13] and the right of access,[14] are also framed in such a way to create obligations on the side of the controller.[15] The controller is in principle also held liable for any damage caused by unlawful processing.[16] All in all, the role of the concept of the data controller is to allocate responsibility in terms of data protection.[17]

It may happen, especially in complicated technical environments, that it is difficult to assess who is acting as a controller[18]. Relevant actors may see themselves as "facilitators" but not as the responsible data controllers.[19] However, the Regulation takes a factual approach providing in Art. 4(7) GDPR that the data controller is "the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data". Determination of the 'means' of the processing does not only refer to technical procedures, but also to the question which data shall be processed, who shall have access to the data or when data are to be deleted—

[12] Art. 89(3) GDPR provides that the EU or Member States can enact exceptions from the rights granted in Arts. 15, 16, 18, 19, 20, 21 GDPR.

[13] Arts. 13 and 14 GDPR.

[14] Art. 15 GDPR.

[15] Art. 29 Working Party, Opinion 1/2010 on the concepts of "controller" and "processor".

[16] Art. 82 GDPR.

[17] Art. 29 Working Party, Opinion 1/2010 on the concepts of "controller" and "processor".

[18] See for a more detailed analysis of the assessment of who is acting as controller in complicated technical environments: Mahieu, R., van Hoboken, J., Asghari, H. (2019). Responsibility for Data Protection in a Networked World—On the Question of the Controller, Effective and Complete Protection and Its Application to Data Access Rights in Europe. JIPITEC, 10(1), 85–105. A critical analysis of Art. 26 GDPR is also provided by Kartheuser I. & Nabulsi S (2018). Abgrenzungsfragen bei gemeinsamen Verantwortlichen—Kritische Analyse der Voraussetzungen nach Art. 26 DS-GVO. MMR 21(11), 717–721

[19] Ibid, p. 11.

which amounts in sum to determining the *why* and *how* of the processing activities.[20]

In the context of SoBigData, where many actors are involved, one of the challenges was to determine who in which scenario is acting as a data controller and to clarify the role of the platform operator. The facilities of the platform include among others a data set and method catalogue which enables users of the platform to search for suitable data sets and methods for their research and a web-based environment for running experiments. However, the actual provision of data sets to final users or the analysis of data sets on request of a user are solely under control by the data set providers, who qualify precisely for this reason as the data controllers in the first place. Therefore, it is for them to act according to the obligations of data controllers under the GDPR. For example, in the event the data controller asked for the informed consent of the data subject as the legal basis of the processing and the data subject withdrew consent subsequently, the data must (unless an exemption applies) be erased in accordance with Art. 17 GDPR. If the consent allowed transferring the data and data had been transferred to third parties before the withdrawal, the controller will need to communicate the erasure to each recipient to whom the data got disclosed according to Art. 19 GDPR. The controller also shall inform the data subject about the recipients on request. Final users also become data controllers when they have downloaded a data set that contains personal data. From that moment on, they also have to comply with the applicable data protection provisions. The platform operator as such is not to be regarded as a data controller with respect to the research data sets. The platform is to be regarded solely as an intermediary enabling researchers to look for appropriate data sets and other means for their research. However, as we will see later (Sect. 3), the SoBigData platform is providing many tools and methods to promote lawful and ethical processing of personal data for research purposes.

2.2 Intellectual property law

In parallel to data protection, the sharing of content via SoBigData RI brings issues of intellectual property rights (IPR) into play. First of all, the data sets, both as methods, applications and content made available for sharing via the SoBigData RI, may be protected by IP rights. Secondly, the use and sharing of such content items may be governed by individual licence terms. For instance, software applications, data sets generated from social media content like Facebook, Twitter, Flickr, such as blogs, commentaries, tweets and photographic works, which constitute original creations of the mind, are protected by copyright, whatever may be the mode or form of expression.[21] The sharing of IP-protected content items, either by digital processing, upload, download, translation, modification, constitutes copyright-relevant actions and requires authorization of the right holder.[22] And this right holder is not necessarily the author.[23] In software applications, the authorization is given by release into RI under individual licence terms. For instance, M-Atlas, a mobility querying and data mining system centred onto the concept of spatio-temporal data, is licensed under Academic Free License 3.0.[24] However, social media content may be governed by either individual licence terms or terms of the network or both. For example, the use of data sets collected from Twitter, in particular, Twitter Dataset 2013–2014,[25] Disease Twitter Dataset[26] are subject to the Terms of Twitter. And according to the Twitter Terms, by submitting, posting or displaying content on Twitter, the user grants Twitter a worldwide full-fledged copyright licence with the right to sublicence[27] and Twitter by releasing the Twitter APIs to developers grants the licence to integrate Twitter content into the services or conduct analysis of such Twitter content, but subject to the Terms of Twitter.[28]

The management of IPR both as sharing of IP-protected materials in a legally compliant way requires legal and technical solutions. The integration of such solutions into SoBigData RI is discussed in the subsequent sections next.

2.3 Ethics

While the GDPR is currently the most comprehensive worldwide attempt to codify digital ethics norms in a legal

[20] Ibid, pp. 11–14.

[21] Article 2 Berne Convention for the Protection of Literary and Artistic Works; Article 1 WIPO Copyright Treaty; Article 9 TRIPS Agreement.

[22] Directive 2001/29/EC of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society, OJ L 167 , 22/06/2001 P. 0010–0019.

[23] According to the work-for-hire doctrine, copyright in computer programs developed in course of employment pass to the employer. See: Article 2(3) Directive 2009/24/EC of 23 April 2009 on the legal protection of computer programs.

[24] M-Atlas. Available at https://sobigdata.d4science.org/group/sobigdata-gateway/data-catalogue.

[25] Twitter Dataset 2013–2014: The data set was collected by the Archive team through the Twitter Streaming API which provides free access to 1% of public tweets. Available at https://sobigdata.d4science.org/group/sobigdata-gateway/data-catalogue.

[26] Disease Twitter Dataset: This Twitter Dataset covers two recent outbreaks: Ebola and Zika. About 60 million tweets were collected through a query-based access to the Twitter Streaming API, covering the period of April 13th 2015 to August 2nd 2016. Available at https://sobigdata.d4science.org/group/sobigdata-gateway/data-catalogue.

[27] Article 3 Twitter Terms of Service. Available at https://twitter.com/en/tos.

[28] Paragraph I B i licence from Twitter, Twitter Developer Agreement, Effective: May 25, 2018. Available at https://developer.twitter.com/en/developer-terms/agreement-and-policy.html.

framework, it remains to be seen whether the desired goals, i.e. the protection of individuals in the age of big data, will actually be achieved. It is only in the practice of the coming few years that it will become clear how much regulatory power the law will have. In general, the law could be said to be an expression of more fundamental ethical values and rights which can perhaps never perfectly be codified in particular determinations.

This is why a research framework, if it desires to be truly responsible, needs to go beyond the required legal minimum. In order to do so, we have tried to ground the framework of SoBigData in an analysis of the fundamental issues we claim are raised by the age of big data. We distinguish here between four major categories of moral issues:

- *Moral autonomy and moral identification* The effects that data gathering activities have on the self-determination of the individual
- *Information-based harm* Direct or indirect harms caused to the individual as a result of data
- *Informational injustice* Issues of contextual integrity, where data leave the original context in which it was granted
- *Inequality caused by information asymmetry* Power imbalances in society and buyer–seller interactions caused or made worse by asymmetries in information.

While these values are the core for what it means to do responsible big data social science research, they generally need to be balanced with other considerations. As social science cannot exist without social information, information about individuals—or that may be virtually connected to individuals—is an asset without which the field could never proceed. That is why dilemmas are a fundamental and unavoidable part of this variety of research; each case involves a balancing act between the ends the research is trying to achieve and the means necessary to reach this end, which will usually include information attributed or attributable to human beings. These dilemmas can only be addressed adequately if this balancing act is always made explicit. We interpret the use specification in the widest ethical sense of the word, in which it requires, on the part of the researcher, an explicit formulation of the specific goal the research is seeking to achieve. While there are also hard ethical and legal boundaries (enshrined in the GDPR, for example), in many cases it is much more permissible to collect a vast quantity of information for a noble purpose than for an ignoble one. To do ethical research is therefore inextricably linked to a conception of the public good, which is ultimately the only thing that may legitimate vast information gathering involved in big data. We desire first and foremost that this trade-off, between means and ends, is made abundantly explicit in the papers and proposals of the researchers involved in this project.

3 From law & ethics to concrete implementations

A research infrastructure on which social data including personal data is processed should provide an implementation of the legal and ethical principles discussed above.

Since the introduction of the GDPR, the regulation includes the explicit obligation to apply data protection by design and by default. However, practical implementation of this requirement is severely lacking behind in practice. First, efforts to implement this requirement are often concentrated on data minimization and data security. This is a remnant of the history of privacy by design, which, when it was first developed focused on these aspects [14]. By building our framework on the broader oriented approach of value-sensitive design, we can overcome this limit. Second, there has been "a failure to communicate clearly and directly with those engaged in the engineering of information systems, and a failure to provide the necessary incentives…"[29] By developing what we call a value-sensitive *institutional* design, we aim to overcome these problems of communication and lack of incentives. We designed tools and an institutional structure that helps users of the platform, who are mostly computer scientists, to understand and comply with their legal obligations and to proceed with their research in an ethical manner. And we took account of and worked on incentive structures of the relevant actors that are conducive to reaching these goals.

Firstly, we consider it is important to give researchers a basic understanding of the legal and ethical requirements of their research. This is why the RI provides a massive open online course (MOOC) and ethics briefs for more specific information on ethical aspects concerning particular data sets (Sect. 3.1). Secondly, terms of use of the SoBigData gateway make the users aware that materials hosted by RI are available under individual licence. The metadata in the data set catalogue also provides important information in that respect for the researchers to consider (Sect. 3.2). Thirdly, in order to help the data controllers to comply with their obligations under the GDPR, a data protection assessment form is made available (Sect. 3.3). Fourthly, in order to create transparency

[29] Bygrave, L. A. (2017). Data Protection by Design and by Default: Deciphering the EU's Legislative Requirements. Oslo Law Review, 4(02), 105–120. https://doi.org/10.18261/issn.2387-3299-2017-02-03. See also Mahieu, R., van Eck, N. J., van Putten, D., & van den Hoven, J. (2018). From dignity to security protocols: a scientometric analysis of digital ethics. Ethics and Information Technology, 20(3), 175–187. https://doi.org/10.1007/s10676-018-9457-5 showing a divide between the work on digital ethics in the fields of ethics, law and computer science.

with regard to the specific research process, researchers are requested to publish information about their research online via a public information document (Sect. 3.4). Furthermore, SoBigData provides a privacy risk assessment methodology for enabling privacy-aware social data analysis and mining (Sect. 3.6). Finally, we established an operational ethical board (Sect. 3.5), which actively participates in research activities.

3.1 Ensuring awareness about legal and ethical requirements among the researchers via a MOOC and ethics briefs

One of the main pillars of the SoBigData legal and ethical framework is creating awareness on side of the users. Knowing the legal risks and ethical issues enables data-driven research based on more awareness about possible ethical consequences. Only by relying on this premise, it is possible to unlock the immense potential of big data analytics for both innovation and social good. To a large extent, however, legal and ethical issues relating to the use of personal data in social science are still a matter for specialists. Ideally, we would like everybody, involved in research based on personal data, to be aware and act with ethical and legal considerations in mind.

In order to achieve this goal, SoBigData provides an online course, which provides the basis of ethical issues involved in the managing of personal data, especially regarding the access and use of the SoBigData RI. The course is accessible (through registration) at the web page http://fair.sobigdata.eu/moodle/. It is organized in three modules covering an overview of the main ethical and legal problems, the definition and the obligations of a data controller and the intellectual property law. The plan is to expand the course offer, adding, for example, modules specific on research purposes and a summary of the main privacy-enhancing technologies. Each module has an associated quiz composed of 6–10 multiple-choice questions. The primary purpose of the quiz is to test and increase the knowledge of researchers. Indeed, there are no particularly tricky questions, but a general competence is required to answer. Moreover, for each (wrong) answer, we provide immediate feedback, which explains why the answer is incorrect and makes a link available to the part of the lesson that debates the topic covered by the question.

Aside from the more general issues tackled by the MOOC, there are also particular problems associated with specific forms of data, which are addressed through "ethics briefs". Handling GPS data, for example, involves issues of coarse-graining spatio-temporal information. Indeed, GPS data record the spatio-temporal movements of people revealing the location visited by an individual and the time of the visit. Knowing the places visited by people together with the time of the visit could enable the disclosure of sensitive information. For example, the systematic visit of a place of worship leads to infer the individual's religion belief. Twitter data instead require careful considerations of the intellectual property rights involved. When legal and ethical issues are presented around concrete research situations, researchers have more incentive to engage with these topics. Abstract ethical issues can be perceived more vividly when presented in the particular domain someone is working in. Moreover, more concrete solutions can be proposed. This is why we deem it is important to produce ethics briefs associated with the virtual environment with particular kinds of data. These briefs inform the researcher of both the problems they may face with a particular kind of data and the solutions that already exist to these problems. They also contain real-life examples of issues that have arisen in big data applications and which users may encounter on the job. Clearly, ethics briefs can be exploited as learning material, which exposes potential trainees to real-life problems and issues they would encounter on the job; they provide them guidelines on how they could manage these problems and concerns, and they taught them principles and practices that apply to specific scenarios. Considering the tremendous variety in data sets, this is still an ongoing process, but we have already produced a number of proofs of concept (i.e. Twitter, DE Webarchive, human mobility data, call detail records). As a practical example, in the case of use of GPS data for smart cities applications, the ethics briefs include the following information and considerations:

- Explanation of privacy threats related to this type of data: for example, a seemingly innocuous data set like taxi fares records could reveal, if it is cross-referenced with other knowledge, personal information on passengers, such as which of passengers were celebrities or visited strip clubs.
- Suggestions on privacy risk assessment methodologies and tools, useful for identifying the users in the data who are at risk: an example of this point will be extensively discussed in Sect. 3.6.
- Suggestions on appropriate measures aimed at minimizing the possibility of (re-)identification of individuals: for example, if the data analysis is about traffic, we do not need to know where a specific person is. Thus, using a methodology to reduce the granularity of a data set, while preserving its salience, might allow for each individual to be protected from unwanted attention [21,22].
- Suggestion on appropriate literature and references to concrete examples.
- Highlights on the importance of the principles of deliberation and openness. This aspect is quite general, as it can be applied to other kinds of data. However, the main points are: get informed about the topic, discuss ethical

deliberations with others, such as the institutional review board of the own institution.
– Information of a contact person from SoBigData, who is an expert in the management of mobility data, and who can help in handling ethical issues related to this data.

3.2 Ensuring compliance with intellectual property law via SoBigData Gateway Terms of Use and metadata

SoBigData integrates a legal and technical solution to manage the IPR issues associated with data sharing via the RI. The legal part is incorporated into the SoBigData Gateway Terms of Use.[30] By accessing the SoBigData RI, which is hosted by D4Science.org Gateway, the users are made aware that materials hosted by RI are available under individual licence terms, indicated in metadata, and agree to follow those terms by using individual items.

The licence terms, basic rights and scope of use associated with individual data sets are indicated in the metadata attached to the data items. For instance, the metadata of Disease Twitter Dataset[31] contains information about accessibility mode (online access); data set availability (onsite, which is in line with the Twitter Developer Policy[32]); scope of use (research only); data source (Twitter); use and distribution requirements (Twitter Terms). The awareness of users about the IPR rules associated with the conduct of research on SoBigData RI is facilitated by an IPR section in the MOOC.

The governance of copyright by the SoBigData Gateway Terms of Use together with the tools for users awareness about the terms associated with the use of individual data items via metadata create a comprehensive framework enabling data-driven research via SoBigData RI in a compliant and respectful way.

3.3 Ensuring accountability via the data protection law assessment form

Data controllers processing personal data for research must comply with the data protection principles, and they have to be able to demonstrate compliance with those principles (Art. 5 (2) GDPR[33]). In order to allow researchers to assess whether their research is compliant with these principles, SoBigData provides a data protection law assessment form.[34] This document is a concise check list of questions regarding important legal requirements to be taken into account in the data-driven research (i.e. data, processes, security and privacy measures). It represents a filter to the relevant aspects in comparison with the legal text of the Regulation or legal textbooks. Its conciseness enables a structured analysis of the legal requirements of a research in a logical order. An associated guide enables the researcher to self-study the various aspects to consider and to have a better understanding of the several legal points mentioned in the list. In order to enhance understandability, the guide also encompasses specific examples and suggestions for further reading. This form is not a simple checklist but, rather, it requires to researchers to give short explanations why they think they fulfil certain legal requirements in order to avoid premature decisions. Despite the described efforts to accommodate the list to non-privacy experts, researchers may still find some of the data protection aspects cumbersome to assess. Too much simplification, however, may result in unintended illegal actions on the side of researchers, which is why we did not go down this path. In the case researchers are not experienced in the field of data protection law and have some difficulties in filling this form, they have the opportunity to contact the SoBigData help desk which can assist researchers: the operational ethical board (Sect. 3.5).

3.4 Ensuring transparency via the public information document

Many of the data practices that are taking place today are opaque. This is also the case for the use of personal data in social science. Transparency is one of the central principles mandated for the processing of personal data by the GDPR (Recital 39 and Art. 5(1)(a) GDPR). There are also comprehensive information duties on part of the data controllers and specific data subject's rights to request information from the data controllers (Art. 15 GDPR). As previously mentioned (see Sect. 2.1.5), the GDPR alleviates some of the transparency burdens, in particular when processing takes place for scientific purposes. However, in these cases Art. 14(5) GDPR requires that information must be made publicly available. Art. 11 GDPR also restricts information obligations and data subject's rights.

In SoBigData, and in any research project in general, in some cases, data subjects cannot be individually informed, because the effort would seriously impair the research project or because the data controller does not have access to the data subjects' contact information. In these cases, data controllers

[30] SoBigData Gateway Terms of Use. Available at https://sobigdata.d4science.org/terms-of-use.

[31] Disease Twitter Dataset accessible via SoBigData Catalogue accessible at: https://sobigdata.d4science.org/catalogue.

[32] Section I.F Be A Good Partner to Twitter, Twitter Developer Policy, Effective: November 3, 2017. Available at https://developer.twitter.com/en/developer-terms/policy.html.

[33] Art. 30 GDPR also requires that each controller shall maintain a record of certain information, e.g. the name and contact details of the controller, its representative and the data protection officer or the purpose of the processing.

[34] This is to be distinguished from the data protection impact assessment according to Art. 35 GDPR.

should make information about their research public because transparency is a precondition for accountability.

By opening up a process, it becomes possible for others to critically assess the process and thus hold accountable those who control it. For this reason, transparency is widely discussed in political philosophy and is a cornerstone of liberal democratic societies. Transparency is also core value for the responsible use of personal data. According to Gürses [12], both EU privacy regulation and the US Fair Information Practice Principles (FIPPs) demand technological and organizational measures that primarily focus on transparency. She claims that the reason for the centrality of transparency is that "if these mechanisms are in place, ideally users can make informed decisions about and have greater control over the collection and flows of their personal information. Further, abuses can be detected or mitigated".

Therefore, one of the principles that SoBigData uses is "Inform the data subjects". In line with the intent of the law, this principle intends to create an environment of openness. The idea is to make SoBigData RI transparent about the processing of personal data that is enabled through the RI itself. In particular, SoBigData provides an environment that is conducive to openness to the public by making public information documents about the databases available in RI. These documents shall enable citizens to obtain information about the personal data that is being used by the researchers in the SoBigData context. As a consequence, for any data set in the RI, researchers must prepare this public document which is written in a simple manner and contains compact information with respect to the research project, the data controller, the steps taken to ensure privacy and data protection and all other information that must be made available according to Art. 14(1) GDPR.

As an example, some of the questions we ask to SoBigData researchers are:

- Which data sets and methods are used in a particular research project or exploratory.
- Which is the pursued research and what is the public benefit of that research.
- What are the potential privacy risks involved, and what are the steps we are taking to minimize these risks to privacy. The security problem is also taken into consideration.
- Where the data originated from, and if the source is publicly available.
- Who is the (representative of the) data controller, and how he/she can be contacted.
- On what legal basis personal data are processed.
- With whom data are shared and who can access the data.
- How long the data are intended to be maintained.
- What procedures are set in place for a citizen to request access, rectification, restriction of the processing.

3.5 Providing support to researchers via the operational ethical board

We deem it important that there is a flexible and responsive part of the infrastructure that can address issues as they may arise. While an ethics board was already instituted, we found that it was populated by senior researchers who (as in most ethics boards) do not have the time to evaluate the numerous individual research projects that go through SoBigData. This is why we have instituted an interdisciplinary (ethics, legal, technical) operational ethical board (OEB), consisting of researchers who are more readily available, and easier to approach. Members of an operational ethics board have more incentive to engage with practical ethical questions in specific research use cases than a high-level ethics board which can only engage in high-level questions. The operational ethical board is equipped with three major tasks:

- Evaluating research proposals and articles on the basis of ethical/juridical standards, as well as offering support in addressing these challenges.
- Expanding and improving on the information architecture (MOOC, ethics briefs).
- Briefing the consortium on the latest developments in digital ethics and law.

The work of the OEB could be replicated in institutions and companies that have access to social data used in analytical processes, with the aim of providing stable support in ethical best practices to data scientists. Moreover, in institutions where an institutional review board (IRB) is present, the OEB could offer a quicker communication channel w.r.t. IRB, at least to provide first advice or opinion, thus reducing the requests to the IRB and relieving IRB of some of its work. Lastly, further communication between data scientists and IRBs could be sensibly improved thanks to the attendance of the MOOC, which can be used as an alignment tool to acquire common knowledge, language, and background between data scientists and members of ethics boards. In particular, the MOOC can be attended by both parties. On the one hand, it provides data scientists to learn basic ethical notions and study practical examples of real cases, useful for starting to face ethical questions. On the other hand, the MOOC provides to IRB's members (who are commonly ethicists) a concrete possibility to be constantly updated about essential technologies and techniques, suitable for a variety of contexts.

3.6 Privacy risk assessment

Whereas the tools and principles are directed at the organizational level and together represent the beginning of a value-sensitive institutional design, technical privacy-by-

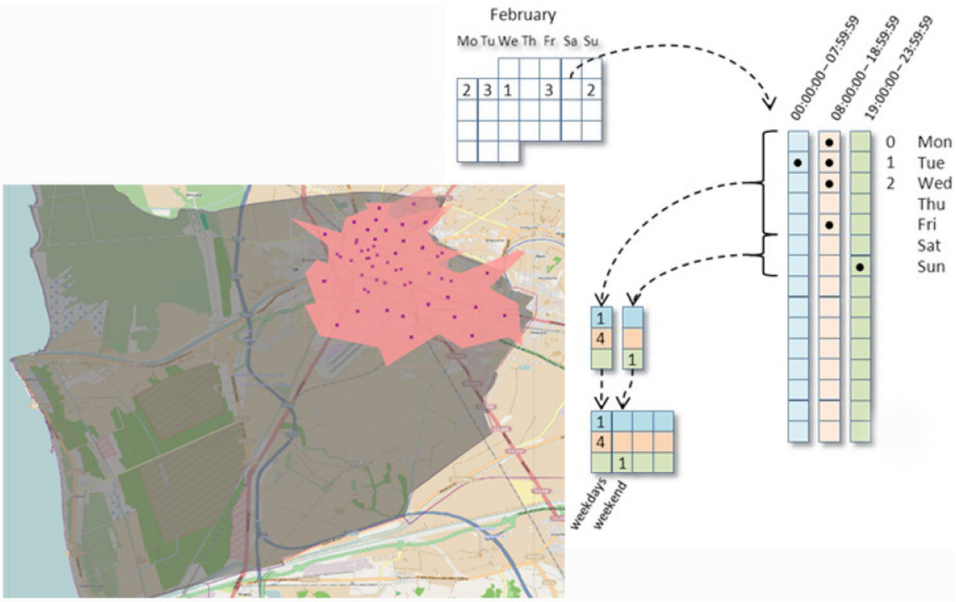

Fig. 1 Methodology for GSM user profile construction: (left) CDR cell coverage in the area of Pisa; (right) schematic representation of reconstruction of the Temporal Profile for users in Pisa. Firstly, starting from the call activities, the presence of each user is registered in the correspondent time slot of each day; then, for each time slot, the number of presences of the user is computed, discriminating between weekdays and weekends

design methodologies are also developed within SoBigData. SoBigData is the first RI to address the privacy issue in a systematic way, providing tools and methodologies that effectively enable both the empirical evaluation of the privacy risk and data transformations by using privacy-preserving approaches available in the infrastructure. To this end, SoBigData provides a set of privacy-by-design methodologies that permit the transformation of personal data into anonymous data for enabling privacy-aware social data analysis and mining. The RI also makes available a privacy risk assessment framework that enables the assessment and the mitigation of the privacy risk, named PRUDEnce [26]. This framework permits the systematic exploration of both: (a) the empirical privacy risks with respect to different attack models and (b) the relationship between data quality and privacy risk level. This leads to the selection of a privacy transformation that is focused only on the risky data identified by the systematic reasoning.

The SoBigData approach is completely different from the typical technical solutions used in the literature for addressing the privacy issues in contexts based on the cloud paradigm. Indeed, in this setting, the most common methodology for approaching the privacy issues is to setup and define access control policies enabling different data views and authorizations for different users. Since, typically, the assumption is that the provider of the (cloud) infrastructure is often not considered fully trusted, then encryption-based approaches are applied [5,6,11,32]. In these approaches, the authorizations are enforced by ensuring that encryption depends on the values of certain attributes characterizing authorized users. This allows the user to access data only if his/her set of attributes matches conditions on the attributes associated with the encrypted data. Other proposals instead consider the possibility of setting up usage control policies that extend standard access control models. This is done by introducing the possibility to evaluate the access authorizations dynamically, thus revoking the permission when conditions change [24,25]. An implementation of usage control policies in cloud computing is presented in [7], where the policies regulate the usage of resources of the cloud infrastructure. The authors describe the integration of a usage control-based authorization service within the cloud service OpenNebula.

3.6.1 Privacy risk assessment for call profiles

In this section, we present a practical application of the SoBigData methodology for the privacy risk assessment on CDR data (i.e. call detail records). CDR is a data record produced by a telephone exchange or other telecommunications equipment that documents the details of a telephone call. This type of data registers the presence of users in a territory enabling a data-driven methodology, called Sociometer, able to classify citizen, based on their call activity profiles, into three categories: residents, commuters and visitors. The call profile is a spatio-temporal aggregation representing the presence of a user in a certain area of interest during different predefined time-slots. The idea is that if a person makes a call in the area A at time t, it means that the user is present in that area at that time.

Figure 1 summarizes the analytical process that leads to the construction of a call profile given the CDR of a user. Once the profiles have been created, the Sociometer classifies them implementing a set of domain rules that describes the mobility behaviour categories. Details on the Sociometer can be found in [10].

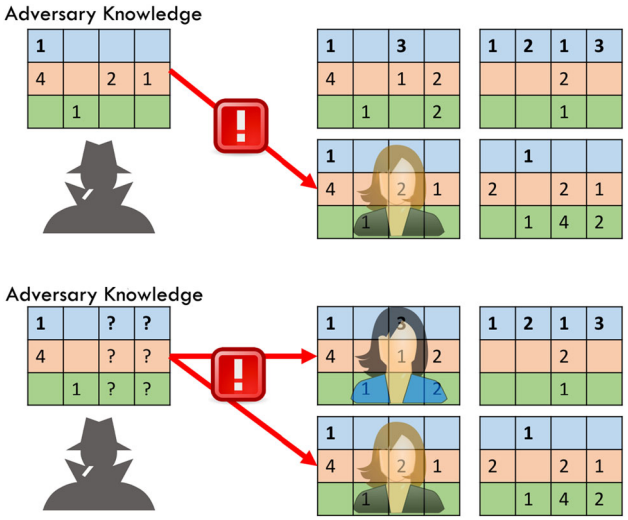

Fig. 2 Example of privacy attacks: (above) The adversary Eve knows all the activities performed by his target, Alice, in 2 weeks; Eve is able to re-identify correctly Alice, (below) Eve only knows 1 week of Alice's activities, so in this case Eve found 2 matching profiles and she is not sure which is the one related to Alice

Call profiles are more aggregated w.r.t. the CDR logs because they cannot reveal the history of the user movements, the number of calls and the exact day and time of each call. Although the only information that one can infer is that a specific user visited a city in a specific aggregated period, this could lead to privacy breaches. As an example, an adversary could understand that a given user went to Pisa during a specific weekend if: (a) the profiles that he was accessing are related to people in Pisa, and (b) he has enough information about the user call activity to allow the partial reconstruction of the profile.

The privacy risk assessment framework PRUDEnce enables the identification of risky profiles and their anonymization before using them in the Sociometer. We evaluate the privacy risk considering various levels of adversary knowledge: we assume that the adversary knows, for a certain municipality, the activities done by a user U, in particular the time of all his calls, for a period of 1, 2, 3 or 4 weeks. This means that, with this knowledge, the adversary can build exactly the same or a partial profile that is created by the telco operator. The adversary model defines how the adversary uses the background knowledge about the user U to identify all the matching profiles. Finally, a simulation of the all possible attacks on the data is performed in order to evaluate the maximum risk of re-identification. Figure 2 shows two cases of re-identification attacks enabled by two different adversary knowledge. In the first case, the risk of re-identification is 100% because only 1 user in the set has a call profile which corresponds to the adversary knowledge. In the second case, instead the privacy risk is 50% because we have two matching profiles.

Simulating the privacy attack for each user in the data, PRUDEnce is able to provide a systematic evaluation of the privacy risk for each user profile. As a consequence, the SoBigData researcher can obtain a overall picture on the distribution of the privacy risk on the available data.

4 Conclusion

Digital ethics is still an incredibly open field. This is also reflected in attempts to implement value-sensitive approaches in this domain, such as in the infrastructure of SoBigData. Whereas the sphere of medical ethics, for example, has a long history of dealing with ethical issues, a similarly solid collection of regulations, jurisprudence, habits and institutions is still only emerging in the case of big data, which renders the measures developed here still more or less experimental.

There are a few issues in particular that need to be addressed in the future:

1. *The relation between public and private sources of information* The data sets in SoBigData are often subject to the rules of the original data collector, such as a private company, simply because there are at present much more data in the private sector. This makes it difficult for SoBigData to introduce its own norms in the way these data are handled.
2. *Openness versus closure* SoBigData does not exercise complete control over the way in which the research data sets are handled through the research infrastructure. In the end, data providers determine themselves whom they want to give access to their data sets and the researchers themselves decide how to conduct their research. As an intermediary, the control that SoBigData does have is only in allowing or barring access to (a part of) the research infrastructure. Yet there is an obvious tension here with the principle of openness that we want to uphold. This means that we have to decide *how and in what manner we want to be* "*open*".
3. *Vagueness inherent to the ethical and legal norms* Content of core concepts of data protection regulation is not settled. For example, the distinction between personal data and non-personal data is contested, as is the distinction between data controller and data processor. Some, such as EDPS' Buttarelli [4], have even suggested that we may have to revisit our conception of personal data in order to find a firm foundation for privacy and data protection in big data societies.
   For SoBigData, this means that it is important to continue to see how the field of digital ethics will develop in the coming years. It may be that it is only in the jurisprudence on the GDPR, for example, that it will become clear whether the measures devised here are adequate.

Above all, we feel that the problems in digital ethics are complex enough that they can only be addressed collaboratively. That is why we welcome all feedback and dialogue coming from colleagues outside SoBigData, who perhaps are struggling with similar issues, in order to together come up with creative solutions to these issues.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Anderson, C.: The Future of High Tech: The Power of a Strong Startup Eco System. Related by Laurens van de Velde/Universiteit van Twente (2016)
2. Boehme-Nesler, V.: Das Ende der Anonymität—Wie Big Data das Datenschutzrecht verändert. DuD **40**(7), 419–423 (2016). https://doi.org/10.1007/s11623-016-0629-3
3. Bretthauer, S.: Compliance-by-design-Anforderungen bei Smart Data. ZD **6**(2), 267–274 (2016)
4. Buttarelli, G.: Opinion 4/2015 Towards a New Digital Ethics—Data, Dignity and Technology (2015). Retrieved from https://edps.europa.eu/sites/edp/files/publication/15-09-11_data_ethics_en.pdf. Accessed on 31 May 2019
5. Capitani, D., di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Over-encryption: management of access control evolution on outsourced data. In: Proceeding of the 33rd International Conference on Very Large Data Bases (VLDB) (2007)
6. Capitani, D., di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Encryption policies for regulating access to outsourced data. ACM Trans. Database Syst. **35**(2), 12 (2010)
7. Carniani, E., D'Arenzo, D., Lazouski, A., Martinelli, F., Mori, P.: Usage control on cloud systems. Future Gener. Comput. Syst. **63**(C), 37–55 (2016). https://doi.org/10.1016/j.future.2016.04.010
8. European Data Protection Supervisor, Opinion 7/2015. Meeting the Challenges of Big Data. https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-11-19_Big_Data_EN.pdf. Accessed on 31 May 2019
9. European Research Area: The Lund Declaration (2009). Retrieved from https://era.gv.at/object/document/130. Accessed on 31 May 2019
10. Furletti, B., Trasarti, R., Cintia, P., Gabrielli, L.: Discovering and understanding city events with big data: the case of Rome. Inf. Multidiscip. Digit. Publ. Inst. **8**(74), 3 (2017)
11. Goyal, V., Pandey, O., Sahai, A., Waters, B.: Attribute-based encryption for fine-grained access control of encrypted data. In: Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS), Alexandria, VA, USA (2006)
12. Gürses, S.: Can you engineer privacy? Commun. ACM **57**(8), 20–23 (2014). https://doi.org/10.1145/2633029
13. Hasan, M., Rundensteiner, E., Agu, E.: Automatic emotion detection in text streams by analyzing Twitter data. Int. J. Data Sci. Anal. **7**, 35 (2019). https://doi.org/10.1007/s41060-018-0096-z
14. Hustinx, P.: Privacy by design: delivering the promises. Identity Inf. Soc. **3**(2), 253–255 (2010). https://doi.org/10.1007/s12394-010-0061-z
15. Inkpen, D., Roche, M., Teisseire, M.: Guest editorial: Special issue on environmental and geospatial data analytics. Int. J. Data Sci. Anal. **5**, 81 (2018). https://doi.org/10.1007/s41060-018-0105-2
16. Krügel, T.: Das personenbezogene Datum nach der DS-GVO—Mehr Klarheit und Rechtssicherheit? ZD **7**(10), 455–460 (2017)
17. Katko, P., Babaei-Beigi, A.: Accountability statt Einwilligung? Führt Big Data zum Paradigmenwechsel im Datenschutz. MMR **17**(6), 360–364 (2014)
18. Marnau, N.: Anonymisierung. Pseudonymisierung und Transparenz für Big Data. DuD **40**(7), 428–433 (2016)
19. Martini, M.: In: B. Paal and D. Pauly (eds.) Datenschutz-Grundverordnung. München: C.H. Beck (2017)
20. Mayer-Schönberger, V., Padova, Y.: Regime change? Enabling big data through Europe's new data protection regulation. Colum. Sci. Tech. L. Rev. **17**, 315–335 (2016)
21. Monreale, A., Rinzivillo, S., Pratesi, F., Giannotti, F., Pedreschi, D.: Privacy-by-design in big data analytics and social mining. EPJ Data Sci. **3**(1), 10 (2014)
22. Monreale, A., Andrienko, G.L., Andrienko, N.V., Giannotti, F., Pedreschi, D., Rinzivillo, S., Wrobel, S.: Movement data anonymity through generalization. Trans. Data Priv. **3**(2), 91–121 (2010)
23. Narayanan, A., Felten, E.W.: No silver bullet: de-identification still doesn't work. White Paper (2014). Retrieved from http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf. Accessed on 31 May 2019
24. Park, J., Sandhu, R.: Towards usage control models: beyond traditional access control. In: Proceedings of the 7-th ACM Symposium on Access Control (2002)
25. Park, J., Sandhu, R.: The UCONABC usage control model. ACM Trans. Inf. Syst. Secur. **7**(1), 128–174 (2004)
26. Pratesi, F., Monreale, A., Trasarti, R., Giannotti, F., Pedreschi, D., Yanagihara, T.: PRUDEnce: a system for assessing privacy risk vs utility in data sharing ecosystems. Trans. Data Priv. **11**, 139–167 (2018)
27. President of the Council of European Union: Rome Declaration on Responsible Research and Innovation in Europe (2014). Retrieved from https://ec.europa.eu/research/swafs/pdf/rome_declaration_RRI_final_21_November.pdf. Accessed on 31 May 2019
28. Rodríguez-González, A., Vakali, A., Mayer, M.A., Okumura, T., Menasalvas-Ruiz, E., Spiliopoulou, M.: Introduction to the special issue on social data analytics in medicine and healthcare. Int. J. Data Sci. Anal. **8**, 325 (2019). https://doi.org/10.1007/s41060-019-00199-9
29. Sarunski, M.: Big Data-Ende der Anonymität? Fragen aus Sicht der Datenschutzaufsichtsbehörde Mecklenburg-Vorpommern. DuD **40**(7), 424–427 (2016). https://doi.org/10.1007/s11623-016-0630-x
30. Schefzig, J.: Big Data = Personal Data? Der Personenbezug von Daten bei Big-Data-Analysen. K&R **19**(12), 772–778 (2014)
31. Van den Hoven, J.: ICT and value sensitive design. In: The Information Society: Innovation, Legitimacy, Ethics and Democracy in Honor of Professor Jacques Berleur SJ, pp. 67–72. Springer, Boston (2007)
32. Waters, B.: Ciphertext-policy attribute-based encryption: an expressive, efficient, and provably secure realization. In: PKC 2011. LNCS, vol. 6571, pp. 53–70. Springer, Heidelberg (2011)
33. Zuboff, S.: Big other: surveillance capitalism and the prospects of an information civilization. J. Inf. Technol. **30**(1), 75–89 (2015)