Access and Reuse of Machine-Generated Data for Scientific Research

Alexandra Giannopoulou*

Abstract

Data driven innovation holds the potential in transforming current business and knowledge discovery models. For this reason, data sharing has become one of the central points of interest for the European Commission towards the creation of a Digital Single Market. The value of automatically generated data, which are collected by Internet-connected objects (IoT), is increasing: from smart houses to wearables, machine-generated data hold significant potential for growth, learning, and problem solving. Facilitating researchers in order to provide access to these types of data implies not only the articulation of existing legal obstacles and of proposed legal solutions but also the understanding of the incentives that motivate the sharing of the data in question. What are the legal tools that researchers can use to gain access and reuse rights in the context of their research?

Keywords: machine-generated data, Internet of Things, scientific research, personal data, GDPR

1 Introduction

When Nicola Tesla was describing the society of the future, he envisioned Earth as 'a huge brain, which in fact it is, all things being particles of a real and rhythmic whole'. In 1990, John Romkey created what is considered among the first Internet of Things[1] devices; he created a toaster that could turn on and off over the Internet. The challenge was part of a conference, which earned its creator a well-earned place among the exhibitors. At around the same time, Neil Gross described a society that 'don(s) electronic skin. It will use the Internet as a scaffold to support and transmit its sensations'.[2]

The world of interconnected things – that is, things that connect to each other and to the environment – is here: from cars to houses and from body sensors to industry applications, data is being produced at an unprecedented daily pace.[3]

The fast accelerating production of data has led to a natural curiosity over its untapped potential by both private and public actors. For example, and as part of the initiative aiming to create a common data space in the EU, the European Commission published two communications related to the building of a European data economy and addressing the issue of growing accumulation of privately held data. The emergence of the open data movement brought forward the idea that open sharing of special categories of data contributes in achieving transparency, accountability, justice, equality and overall better democratic processes. Consequently, data sharing has stayed at the forefront of several policy proposals and legislative reforms in the latest years. Open government data, open research data, open science and more have all been developed to address social problems through advancements in collecting, accessing, analysing and processing big data. The innovation potential that drives the enhancement of data access and reusability practices illustrates the significant value derived from the expansion of data sharing practices.[4] The transformative effect from the use of data towards serving the goals of a democratic society can be witnessed in our economy and also in research and knowledge production; in fact, knowledge derived from data-based services has the potential to revolutionise citizens' quality of life, to establish the ground that would provide evidence-based policy actions and to create new growth business opportunities. Examples of recent reforms that address the free flow of data on a European level include the General Data Protection Regulation[5] whose goal is to create a normative framework for the free circulation

* Institute for Information Law (IViR) – University of Amsterdam.

1. According to Art. 29 Working Party, it is 'an infrastructure in which billions of sensors embedded in common, everyday devices […] are designed to record, process, store and transfer data and […] interact with other devices or systems using networking capabilities': Article 29 Working Party, 'Opinion 8/2014 on the on Recent Developments on the Internet of Things', *Opinion WP 223*, available at: http://ec.europa.eu/justice/article-29/documentation (last visited 15 April 2019). According to the Federal Trade Commission: 'The Internet of Things ("IoT") refers to the ability of everyday objects to connect to the Internet and to send and receive data'. Federal Trade Commission, 'Internet of Things – Privacy & Security in a Connected World', *FTC Staff Report* (2015), available at: www.ftc.gov/system/files/documents/reports/ (last visited 15 April 2019).
2. N. Gross, 'The Earth Will Don an Electric Skin', *Bloomberg*, 30 August 1999, available at: https://www.bloomberg.com/news/articles/1999-08-29/14-the-earth-will-don-an-electronic-skin (last visited 15 April 2019).
3. It was in 2005 when Jonathan Zittrain described the impact of cheap sensors in augmenting data production and surveillance states: J Zittrain, *The Future of the Internet-And How to Stop It*, Yale University Press, New Haven (2008) 205.
4. OECD, *Data-Driven Innovation: Big Data for Growth and Well-Being* (2015) 195, available at: www.oecd.org/sti/data-driven-innovation-9789264229358-en.htm (last visited 15 April 2019).
5. 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 (hereinafter GDPR).

of personal data as well as the Public Sector Information directive[6] aiming to facilitate participatory democracy, to improve administrative efficiency and to promote economic development through open access to public sector data. Finally, the legal framework is completed by sector-specific legislation among different types of data production, standardised licenses and data policies that regulate and promote special cases of data sharing.

A new data category consisting of data generated by machines and sensors has emerged, qualified under the generic term of 'machine-generated data'. This nascent category has been progressively attracting the attention of both the market and regulation as data is produced en masse from private entities. Machine-generated data is comprised of data automatically generated by a computer process, application or other mechanism without the active intervention of a human.[7] The most prominent examples come from the Internet of Things, whose business model is founded on the automated collection of user data towards ameliorating user experience and services provided. In fact, various business sectors (*i.e.*, the motor vehicle sector with the emerging autonomous car technologies) have already seen significant disruption from the amount of data produced, collected and processed. For instance, the functioning of autonomous cars is largely interconnected with data collection and processing in order to not only perform its basic function but also to provide better services.[8] Similarly, smart homes are comprised of a set of Internet-connected and interconnected devices collecting and processing data in order to produce services that allow for maximum comfort and efficiency. Also, smart thermostats allow energy-saving both through remote controlling of the temperature and through learning the owners' schedules and behaviour. In the agricultural sector, smart farming devices have revolutionised production and the overall economy by permitting the collection, processing and dissemination of data related to the farming processes. The data collected create the necessary breeding ground for the optimisation of farming practices and of energy and overall financial costs. Overall, there is high value and market potential surrounding this type of data, which 'is a primary resource, asset, and product of the digital economy'.[9]

The fast-paced technological environment that relies on the generation, collection and processing of machine-generated data has highlighted regulatory gaps in the process of fostering a data-based economy. In fact, machine-generated data hold two unique traits that should be taken into inherent consideration before implementing any regulatory framework: firstly, it consists of data which are not directly produced by humans and as such they do not automatically fall under the same logics and conditions; secondly, this type of data is predominantly held by private companies that create the objects destined to function in the hands of data-producing users. The companies in question contractually restrict access to data for market purposes. Realising the market value and innovation potential, the European legislator picked up the regulatory challenge for the further fostering of a Digital Single Market.[10] In that context, the European Commission pledged to build a 'data economy'. More specifically, when the 'Free Flow of Data' initiative was announced in 2016, it promised to address the obstacles in the free movement of data. The objective was to establish a regulatory framework on the cross-border use of data especially in the context of the Internet of Things. Similarly, the OECD report on data-driven innovation has also highlighted the importance of sharing big data due to their overall beneficial effect in society.[11] With data science technologies making rapid advances, 'access to data and to the information based on it becomes a strategic and valuable asset'.[12]

The European Commission's guide on the free flow of data[13] points out that innovation based on privately held machine-generated data is lacking because the actors involved in this sector of the data economy do not possess the tools necessary to explore the full potential of the data in question. The 'Guidance on Sharing Private Sector Data in the European Economy' document is one of its kind, which sets the way forward for the regulation of data sharing of private sector data.[14] In fact, data control strategies further raise the entry barrier for new or smaller actors in the current competitive environment related to innovation services. Data sharing is an indispensable tool for innovation in this context, but the proper incentives for sharing actions initiated by private actors are still lacking.[15] The European Commission has recognised that horizontal legislative efforts would be inadequate and premature in the context of machine-

6. Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the reuse of public sector information amended by Directive.
7. The term is not new. Talking about data control systems, Chorafas uses the term machine-generated data to point out that 'information has the amazing ability to generate more information'. D.N. Chorafas, *Control Systems Functions and Programming Approaches*, Volume B, Applications, Academic Press, New York and London (1966) 114.
8. A connected smart car, for example, is susceptible to produce up to 25 GB of data every driving hour.
9. N. Elkin-Koren and M.S. Gal, 'The Chilling Effect of Governance-by-Data on Data Markets', 86(2) *University of Chicago Law Review* 403 (2019).
10. The creation of a Digital Single Market was part of the current European Commission's 'priority projects'. *See* European Commission, 'A Digital Single Market for Europe: Commission sets out 16 initiatives to make it happen', Press Release of 6 May 2015, available at: http://europa.eu/rapid/press-release_IP-15-4919_en.htm (last visited 15 April 2019).
11. OECD, above at n. 4.
12. D.L. Rubinfeld and M.S. Gal, 'Access Barriers to Big Data', 59 *Arizona Law Review* 339 (2017), at 363-64.
13. Commission staff working document, Guidance on sharing private sector data in the European data economy, 25 April 2018, SWD(2018) 125 final.
14. B. Gonzalez Otero, Evaluating the EC Private Data Sharing Principles: Setting a Mantra for Artificial Intelligence Nirvana, 10 *JIPITEC* 66 (2019).
15. Incentives for sharing vary from reputation gains to economic benefits and also to further market prospects. *See* Y. Lev-Aretz, 'Data Philanthropy', *The Hastings Law Journal* (forthcoming 2019), available at: http://dx.doi.org/10.2139/ssrn.3320798 (last visited 15 April 2019).

generated data. The suggestion to create a sui generis ownership right for data producers was met with criticism by academics, market players and civil society. It has become clear that a 'one-size-fits-all' approach cannot be easily envisaged because it is very hard to identify patterns across different types of machine-generated data and across different sectors.

Scientific research is progressively becoming more data-intensive with complex structures of scientific discovery. The development of models based on the abundance of diverse data sets and on the advancement of computer analytic processes such as machine learning have altered the landscape of scientific method. Advances in data analytics, along with the dominance of big data, have revolutionised research even in disciplines that were not fundamentally founded on traditional data collection and analysis. Both the business sector and the European Commission have pointed out the innovation potential and the social benefits that can be drawn from the scientific outputs of the processing of machine-generated data. For this reason, the Commission examined the possibility of creating regulatory pathways for the 'enhanced access to commercially-held data for scientific researchers funded from public resources'. In that sense, the question emerging is whether the machine-generated data market can be regulated by exceptional legal rules that allocate specific access rights to expressly designated actors in order to foster innovation and knowledge production for the broader public good.

The issue raised can be further positioned within the question of data regulation, which dominates the legal discourse under multiple facets. It has progressively been approached through different legal perspectives and disciplines, expressed through the issue of establishing property rights on (personal) data and of regulating access through private ordering mechanisms or through direct sector-specific regulation. This article takes the approach of focusing on a specific category of data, that of machine-generated non-personal data, and aims to evaluate how can access to this specific subcategory data be regulated in order to benefit scientific research.

2 Defining Machine-Generated Non-personal Data

Machine-generated data acquired significant market value due to some of its distinguishing features and also due to its societal impact. In fact, the *volume* of data generated by different sensors, Things or 'machines', in general, is a distinguishing characteristic, since machine-generated data is placed in the broader, encompassing category of Big Data. In addition, the overall demonstrated *quality* of the data sets created by the collected data contributes to the production of valuable insights as an (un)expected outcome. The innovations behind these data processing activities have become common ground for all businesses operating towards offering services for citizens. In this context, data sharing becomes a significant vector for generating innovation and economic growth. On a practical level, many private actors holding large and diverse data sets establish data sharing practices and standardised agreements in order to extract the maximum value from the decision-making and analysis processes. A priori, a universal approach that delineates the context of data sharing has not yet been identified, not only due to the nature and diversity of data produced and shared but also due to the uses related to data sharing and the different actors involved in the process. For instance, data sharing activities that involve sharing of sensitive anonymised data are subject to stricter regulatory regimes[16] than that of other types of non-sensitive data, such as meteorological data.

Admitting the nuances and the diversity in existing data in the domain of machine-generated data, the European Commission has been progressively showing a special interest on 'machine-generated non-personal data'. In fact, while personal data flows are predominantly governed by the GDPR, the Commission recognised the potential in regulating the concave space of non-personal data left by the convex scope of application of personal data regulation. The first issue highlighted even before envisaging the framing of the aforementioned category is defining the non-personal data that would fall under its scope of application. On a fundamental level, the concept of personal data is surrounded by ambiguity. Data categorisation is challenging for many reasons, the major issue being the lack of a clearly delineated definition of 'personal data'. In this context, and even if the qualification as machine-generated is relatively straightforward, this is not the case with non-personal data because it is highly dependent on the personal data demarcation. Without such a demarcation it is impossible to create a trustworthy reliant framework on which to base data sharing practices.

According to Article 4(1) GDPR, the definition of personal data is as follows: 'any information relating to an identified or identifiable natural person ("data subject")'. In a 2007 Opinion,[17] the Article 29 Working Group (A29WP) elaborated on the different components of this definition in order to guide the scope of application and its enforcement by courts. Within the constituting elements of the above definition, lies also the context-dependent approach that characterises data protection regulation. More specifically, the concepts of 'relating to' and 'identified or identifiable' are increasingly volatile and ultimately encompass a broad range of

16. *See*, Art. 89 GDPR.
17. Art. 29 Working Party, 'Opinion 4/2007 on the Concept of Personal Data', *WP 136*, available at: http://ec.europa.eu/justice/article-29/documentation (last visited 15 April 2019).

doi: 10.5553/ELR.000136 - ELR November 2019 | No. 2

data. For this reason, the concept of personal data – as it is currently outlined and enforced – has been criticised for being too broad and inapplicable.[18] The *identifiability* test carries a lot of nuance, as it is further distinguished into directly identifiable and indirectly identifiable data.[19] It is further determined by the A29WP that this test is dynamic, leaving ample room for a more flexible application according to a wide range of factors related to the data in question.[20] A fortiori, the concept of non-personal data incorporates the same inherent fluidity found in the concept of personal data. The two types of data that are included in the category of non-personal data are anonymised personal data (which – due to their nature – escape the GDPR scope of application) and non-identified or identifiable data.[21] It is not within the scope of this article to discuss the evolution of the contextual concept of personal data.[22] It suffices to point out at this stage that the creation of a distinct category for data that do not qualify as personal data as a concave definition to the convex one of personal data is inefficient because of the fluidity involved in personal data qualification. According to Graef *et al.*:

> On the basis of such a contingent and context-based application of the definition of personal data, it is difficult to see how a legislative proposal targeting non-personal data could be applied in practice. We foresee substantial difficulties maintaining two separate legal frameworks, one regulating personal data and another one regulating non-personal data, when personal data cannot be clearly distinguished from non-personal data.[23]

Within this context of increasing 'technological capacities to combine and interpret data, personal data will show up ever more frequently in the zettabytes of the twenty-first century information flows'.[24] Many data protection scholars develop critical approaches of the personal data protection, claiming that the distinction becomes meaningless[25] and anonymity is considered no longer possible.[26] Besides the critical view of the 'all-encompassing notion' of personal data, this broad definition is 'welcomed in light of the aim of data protection law to ensure effective and complete protection of data subjects'.[27] However, at the same time, accepting the duality of personal and non-personal data is at odds with the coming technological reality of constant automated collection and processing of data because in this reality 'any information has the potential to affect people'.[28] According to Koops, it would be more useful for data protection if 'instead of trying fitfully to establish where the border lies between personal and non-personal data, we would allow for categories of data that have certain effects on people when they are processed, regardless of whether or not they relate to identifiable individuals'.[29] Sector-specific regulation for data is an approach that has been proposed by scholars, as a way out of the dissonance created between the innovation potential and the regulatory and market complexities.

Without prejudice to the scope of application of personal data regulation, the scope of the article extends to examining the processing of machine-generated non-personal data for research and scientific purposes. As a matter of fact, according to the GDPR, personal data can be processed for the purposes of scientific research as long as the principle of data minimisation is respected, and based on one of the lawful grounds for processing of Article 6(1). In that sense, the Regulation envisages the implementation of legal and technical protection measures such as pseudonymisation and – when possible – anonymisation of personal data. Pseudonymous data fall under the scope of application of the GDPR, while anonymous data are not subject to the Regulation. According to Article 89, paragraphs (1) and Article 89(2), the processing of personal data for scientific or historical research purposes or statistical purposes can also result in the limitation of data subjects' rights in order to satisfy the purposes of the research in question. Thus, data protection regulation leaves some room for derogations from the absolute protection of the individual control of personal data if these rights risk to 'seriously impair or render impossible the achievement of the research'.[30] Recognising the importance of its potential benefits, the Regulation expressly clarifies that 'the processing of personal data for scientific research purposes should be interpreted in a broad manner

18. For an overview of the positions in favour and contra the current state of the concept of personal data, *see, e.g.* F. Zuiderveen Borgesius, 'Singling Out People Without Knowing Their Names – Behavioural Targeting, Pseudonymous Data, and the New Data Protection', 32 *Computer Law & Security Review* 256 (2016) 271.
19. *See*, A29WP 2007 Opinion on the concept of personal data, above at n. 13.
20. *See, for instance*, the guidelines derived from Rec 26 of the GDPR.
21. According to A29WP's opinion on the concept of personal data, a further distinction can be made between directly and indirectly identifiable data. This distinction serves to underline the context-specific nature of personal data: directly identifiable data are the ones that achieve to single out directly an individual through a specific piece of information and indirectly identifiable data are the ones that single out but through the combination of different data points provided.
22. Established CJEU case law illustrates the application of the context-specific character of personal data. *See, e.g.* Breyer (2016) CJEU C-582/14; *Scarlet Extended* (2011) CJEU C-70/10; Lindqvist (2003) CJUE C-101/01.
23. I. Graef *et al*., 'Feedback from Tilburg University on the European Commission's Proposal', available at: https://ec.europa.eu/info/law/better-regulation/initiatives/com-2017-495/feedback/F8922_en (last visited 15 April 2019).
24. B.J. Koops, 'The Trouble with European Data Protection Law', 4(4) *International Data Privacy Law* 250 (2014).
25. O. Tene and J. Polonetsky, 'Big Data for All: Privacy and User Control in the Age of Analytics', 11(5) *Northwestern Journal of Technology and Intellectual Property* 258 (2013).
26. P. Schwartz and D. Solove, 'The PII Problem: Privacy and a New Concept of Personally Identifiable Information', 86 *New York University Law Review* 1814 (2011).
27. N. Purtova, 'The Law of Everything: Broad Concept of Personal Data and Future of EU Data Protection Law', 10(1) *Law, Innovation and Technology* 40 (2018).
28. *Ibid*.
29. Koops, above at n. 24.
30. European Union Agency for Fundamental Rights, *Handbook on European Data Protection Law*, Publications office of the European Union, Luxembourg (2018) 340.

including for example technological development and demonstration, fundamental research, applied research and privately funded research'.[31] In a similar approach to balancing the benefits of research with those of effective data protection and because the specific delimitation ex ante of the purposes of the processing for scientific research can be quite complex, the GDPR creates a derogation from the requirement of purpose limitation when asking for express consent for research and scientific purposes. The creation of a specific favourable regime towards fostering scientific research and innovation takes into account the fact that the analysis still constitutes personal data processing and should thus be subject to appropriate safeguards in order to ensure a responsible and lawful treatment of personal data.[32]

Frequently, when significant effort is required for de-identification, machine-generated data produced from Internet of Things devices is susceptible to qualify as 'high-dimensional data'.[33] According to Narayanan and Felten, 'high-dimensional data is now the norm, not the exception.…[T]hese days it is rare for useful, interesting datasets to be low-dimensional'.[34] In the case of Internet of Things, personal data is produced, collected and kept in privately owned databases with the consent of the users,[35] and they can be made available to researchers under the conditions set out by the GDPR. However, there is a significant part of machine-generated data that can be qualified as non-personal because of the nature of the data in question or because of the context in which it is processed. The value that can be derived from the raw data generated towards promoting researchers is recognised as a significant scientific tool.[36] Although the GDPR sets a framework with specific conditions applicable to the processing of personal data for research purposes, the conditions under which researchers can access non-personal data remain unclear or subject to access contractual conditions set out by the big data-holder companies. This regulatory uncertainty and the adoption of a very fragmented approach in accessing these data sets limits research and restricts the scientific output of researchers according to the chosen data sharing practices by private entities.

3 Non-regulatory Data Sharing Practices

In 2014, Intel decided to share data sets from smart farming data produced by its agricultural sensors with the researchers from the University of California.[37] More recently, the development of apps during a hacking competition was made possible with the use of data shared by private companies related to smart agriculture, such as Agrisyst, ForFarmers and Hendrix Genetics. Multiple examples of voluntary sharing of data from private companies can be found more and more frequently. Data sharing constitutes an established practice among businesses and private entities in general or between data-holder companies, on the one hand, and institutions or researchers, on the other hand. The sharing of private sector data is not a completely new practice, and it has existed under various denominations such as data philanthropy, data donorships and data partnerships, corporate social responsibility, data collaborativism and 'data for good' . The developed practice includes various aspects of making data available for third actors: it can concern making available privately held data for purposes related to the public sector and the public interest; it can relate to data shared between companies, but it also covers the making available of these data to researchers for purposes of scientific advancement. There is little standardisation in the practice of data sharing principles; freedom of contract prevails and the sharing terms vary depending on the actors involved, the type of the data and the nature of the envisaged uses. As previously mentioned, sector-specific approaches are dominating the market:

> Sector-specific regulation appears as the road to take, since the security interests of the state will most likely need different rules than the prevention of infectious diseases, the protection of the environment or the functioning of smart cities or traffic control systems.[38]

Lacking a specific legal framework, the conditions for granting access and use rights to the machine-generated non-personal data are generally established by the man-

31. Rec 159, GDPR.
32. There have been examples where research results have been published without taking the appropriate precautions against the identification of data subjects: W. Hartzog, 'There Is No Such Thing as "Public" Data', *Slate*, 19 May 2016, available at: https://slate.com/technology/2016/05/okcupids-data-leak-shows-theres-no-such-thing-as-public-data.html (last visited 15 April 2019).
33. According to Cavoukian and Castro, high-dimensional data 'consist of numerous data points about each individual, enough that every individual's record is likely to be unique, and not even similar to other records': A. Cavoukian and D. Castro, 'Big Data and Innovation, Setting the Record Straight: De-identification Does Work', 16 June 2014, available at: www2.itif.org/2014-big-data-deidentification.pdf (last visited 15 April 2019).
34. A. Narayanan and E.W. Felten, 'No Silver Bullet: De-identification Still Doesn't Work', 9 July 2014, available at: http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf (last visited 15 April 2019).
35. Consent is expressed by accepting privacy policies and the terms and conditions – necessary precondition to use the device in question.
36. *See, for instance*, P. Rubens, 'Can Big Data Crunching Help Feed the World?', *BBC News*, 11 March 2014, available at: https://www.bbc.com/news/business-26424338 Last visited 15 April 2019; R. K. Perrons and J.W. Jensen, 'Data as an asset: What the oil and gas sector can learn from other industries about big data', 81 *Energy Policy* 117, (2015).
37. L. Gilpin, "How Intel is using IoT and big data to improve food and water security", 13 June 2014, https://www.techrepublic.com/article/how-intel-is-using-iot-and-big-data-to-improve-food-and-water-security (last visited 15 April 2019); *see also*, Y. Lev Aretz, 'Data Philanthropy', *The Hastings Law Journal* (forthcoming 2019), available at: http://dx.doi.org/10.2139/ssrn.3320798.
38. J. Drexl, 'Designing Competitive Markets for Industrial Data – Between Propertisation and Access', 8(4) *JIPITEC* 257 (2017).

ufacturer of the smart object that generated the data through its use by the user. In these cases, the collector of the data is responsible for the data sharing practices enforced through broadly defined contractual agreements.[39] Economic benefits and reputation gains in performing data exchanges between private actors as well as overall financial incentives have been the main factor in establishing data sharing practices.[40] At the same time, advancements in artificial intelligence technology brought forward the need to train the corresponding algorithms, elevating the value of big data sets such as the ones generated by the Internet of Things. In fact, the large data sets produced are ideal candidates for training powerful algorithms. To this day, the industry stresses that the implementation of current business models involving data sharing practices is possible by relying solely on contract law[41] because it allows for the modularity needed in providing dynamic access and usage rights depending on the nature of the data set and the purpose of the use.

The exercise of access and usage control over the data in question through private ordering has demonstrated that contract law serves as a strong instrument imposing restrictions over the subject matter, potentially even stronger than any legally recognised property right. At the same time, the current landscape of data sharing practices illustrates how the absence of legally recognised economic property rights over such data and databases is not prejudicial to the development of a data-driven economy. In fact, the proposal advanced by the Commission to introduce property rights in data has been met with large criticism by scholars[42] and with scepticism by the industry. The market value of the data generated and its central role in the development of the current economic models is undeniable. However, maintaining access to data for the benefit of the public good in view of its societal value is taken into consideration when designing data sharing principles and when considering regulation.

The type of the data generated and their relevance in the context of a general societal or scientific purpose is inciting private companies to engage in sharing of data in the context of 'data for good' movements or data philanthropy in general.[43] In its guide for sharing private sector data, the European Commission uses the term 'data donorship' to describe the voluntary sharing of private sector data with the public sector. These terms aim to describe an aspect of what is called 'corporate social responsibility'; the term is not new and multiple definitions have been advanced in the last years.[44] According to a renewed strategy of the European Commission,[45] corporate social responsibility implies the obligation of companies to

> have in place a process to integrate social, environmental, ethical, human rights and consumer concerns into their business operations and core strategy in close collaboration with their stakeholders, with the aim of – maximising the creation of shared value for their owners/shareholders and for their other stakeholders and society at large – identifying, preventing and mitigating their possible adverse impacts.

The main elements of the definition of social responsibility are therefore that it is voluntary and that it is found in private entities pursuing public interest objectives that go beyond the pursuit of their private interests and of the compliance with current regulatory and contractual obligations. The potential high relevance of certain types of privately held data towards fulfilling greater societal goals has been recently admitted. According to the Commission, the use of the aforementioned data 'can, for example, lead to a more targeted response to epidemics, better urban planning, improved road safety and traffic management, as well as better environmental protection, market monitoring or consumer protection'.[46] Thus, the concept of corporate social responsibility is undergoing a transformation that aims to incorporate those companies that hold useful for the societal good data. Naturally, facilitating data sharing through collaboration between private actors who hold data considered valuable and interested third parties could result in generating value towards the greater public good.[47]

Without a proper normative framework, private voluntary initiatives have emerged so as to foster the sharing of data across companies, sectors, projects and research-

39. V. Mayer-Schönberger and Y. Padova, 'Regime Change? Enabling Big Data through Europe's New Data Protection Regulation', 17 *Columbia Science and Technology Law Review* 315 (2016).
40. T. Klein and S. Verhulst, 'Access to New Data Sources for Statistics: Business Models and Incentives for the Corporate Sector', Discussion Paper No. 10 (2017), available at: http://dx.doi.org/10.2139/ssrn.3141446 (last visited 15 April 2019).
41. Drexl, above at n. 38.
42. *Ibid.*; B. Hugenholtz, 'Data Property: Unwelcome Guest in the House of IP', Paper presented at *Trading Data in the Digital Economy: Legal Concepts and Tools*, Münster, Germany (2017).
43. As it is explained by Lev-Aretz, the data-for-good movement promotes 'data-driven projects that can increase the efficiency of social initiatives, extend their reach, and better tailor them to specific communities. The data-for-good movement has spotlighted the imperative role of the private sector in producing useful data for social action, sparking an active conversation about models and incentives for sharing. As part of this conversation, the term "data philanthropy" was born': Lev Aretz, above at n. 15.
44. Up to twenty competing definitions of corporate social responsibility have been found: A.B. Carroll, 'Corporate Social Responsibility: Evolution of a Definitional Construct', 38(3) *Business & Society* 268 (1999).
45. European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. 'A Renewed EU Strategy 2011-14 for Corporate Social Responsibility', COM(2011) 681, 25 October 2011.
46. European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. 'Towards a Common European Data Space', *COM(2018) 232*, 25 April 2018 at 12.
47. *See, e.g.* Liander, an energy network administrator in the Netherlands has made data related to energy consumption available in order to permit research and innovation on energy conservation and smart energy use. *See,* F. Welle Donker, B. Van Loenen & A.K. Bregt, 'Open Data and Beyond', 5(4) *ISPRS International Journal of Geo-Information* 48 (2016).

ers. The goal of these initiatives is to create an environment encouraging contributions and sharing through the use of contractual tools. The agreements in question are generally described by the term 'data collaboratives',[48] which refers 'to a new form of collaboration, beyond the public-private partnership model, in which participants from different sectors — including private companies, research institutions, and government agencies — can exchange data to help solve public problems'. They are voluntary initiatives created to facilitate access to various types of data and for different uses or for the benefit of different actors.[49] The term primarily used by Stefaan Verhulst and David Sangokoya[50] is not devoid of criticism. As it is described by Yafit Lev-Aretz,

> the term data collaborative is both under-inclusive and over-inclusive. The emphasis on collaboration leaves many instances of data sharing outside the scope of data collaboratives. For example, open data initiatives in the private sector, where datasets are released to the public with no continuous interaction between the public and the provider of the data following the release, can hardly be described as collaborative. The data collaboratives universe (…) does not underscore the sharing of privately-held data or privately owned data-driven insights. It fails to highlight the monetary and business value of the data and does not reflect the ecosystem in which private sector data is shared.[51]

The author uses the broader term of 'data philanthropy', which she defines as being the 'combination of three elements: (1) unpaid for sharing of or access to (2) privately held data or proprietary data insights for (3) the greater good'.[52]

A lot of predominant data sharing practices that subsequently result in the creation of data collaboratives aim towards advancing research with the goal of deriving knowledge from the large amount of existing data. While there is not a distinct procedure to facilitate the sharing of machine-generated data, multiple initiatives of data sharing for research have emerged lately,[53] not without inciting controversy over risks related both to data protection violations[54] and to the lack of informed consent from data subjects as to the further processing of their personal data. Current examples that illustrate the market potential of sharing privately held machine-generated data (both personal and non-personal) for the purposes of advancing academic research also showcase the absence of standardised approaches and the lack of legal clarity in the enforcement of rules in order to make the data economy work.

4 Normative Framework in Extracting Knowledge from Data

The amount of data generated, processed and generally controlled by the industry as well as its prospects as a precious tool for data-driven services has not gone unnoticed from the legislator on a European level. Firstly, the predominant normative tool for data sharing is data protection and privacy regulation. The European Union has created a solid framework for producing digital trust, a precondition for the sustainable development of the data economy. According to the European Data Protection Regulation (GDPR), which entered into force on 25 May 2018 replacing Directive 95/46/EC, 'natural persons should have control of their own personal data'. In that sense, the GDPR guarantees individual autonomy and attributes rights to data subjects that would prevent non-intended uses of their personal data. The European data protection regulation aims to enforce a balancing act between protection and free flow of personal data in order to protect the individual rights without stifling economic potential of data. In order to lay the foundations for a future competitive advantage and according to the Commission's plans to create a European harmonised data-based digital economy, the regulation of the free flow of data within the EU implies the regulation of personal data and the restriction flows that lie beyond this type of data. The main solutions

48. S. Verhulst and D. Sangokoya, 'Data Collaboratives: Exchanging Data to Improve People's Lives', *Medium*, 22 April 2015, available at: https://medium.com/@sverhulst/data-collaboratives-exchanging-data-to-improve-people-s-lives-d0fcfc1bdd9a (last visited 15 April 2019).
49. For a proposed taxonomy of data collaboratives, *see,* I. Susha, M. Janssen, S. Verhuulst, 'Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development', in *Proceedings of the 50th Hawaii International Conference on System Sciences* (2017) 2691.
50. *Ibid*.
51. Lev-Aretz, above at n. 15.
52. *Ibid*.
53. The latest example can be found in the Social Science One project. It consists of a specifically designated expert commission responsible for handling access to data from Facebook for research and scientific purposes. According to the commission in question, 'Social Science One implements a new type of partnership between academic researchers and private industry to advance the goals of social science in understanding and solving society's greatest challenges. The partnership enables academics to analyse the increasingly rich troves of information amassed by private industry in responsible and socially beneficial ways. It ensures the public maintains privacy while gaining societal value from scholarly research. And it enables firms to enlist the scientific community to help them produce social good, while protecting their competitive positions'. The first thematic area will be focused on projects related to 'the effects of social media on democracy and elections', available at: https://socialscience.one/.
54. For instance, the data sharing deal between Google's DeepMind and Royal Free NHS Foundation Trust was determined to be violating of data subjects' privacy according to the ruling issued by the Information Commissionaire's Office (ICO) in the United Kingdom. According to Elizabeth Denham, Information Commissioner, 'there's no doubt the huge potential that creative use of data could have on patient care and clinical improvements, but the price of innovation does not need to be the erosion of fundamental privacy rights'. *See*, ICO's letter outlining the results of the investigation: E. Denham, RFA0627721 – provision of patient data to DeepMind, 3 July 2017, available at: https://ico.org.uk/media/action-weve-taken/undertakings/2014353/undertaking-cover-letter-revised-04072017-to-first-person.pdf (last visited 15 April 2019).

raised surrounding ownership and access rights on non-personal data did not finally manage to produce a legal framework as intended by the Commission's 'Free Flow of Data' initiative announced in 2016.

Data access regulation takes multiple forms and can be found in different normative approaches. For instance, the latest example comes from regulation regarding data mining processes. Data mining has been one of the core issues at the data flow agenda of legislative efforts both at the European and national levels. As a matter of fact, the public interest in allowing text and data mining for (at least) research purposes – if not for all purposes – is gradually being recognised for its societal and economic benefits. It is considered to be a fundamental tool for researchers of all disciplines.[55] Data mining refers to an ensemble of computer science techniques used to extract knowledge from large digital data sets, by looking patterns that are usually difficult to notice with human only research. Data mining is a subset of 'knowledge discovery in databases'. While it may not be perfect, the mining analogy serves to explain roughly what content mining entails. Machine learning algorithms go through large amounts of data, eventually finding valuable information and gaining insights by making combinations that were difficult to foresee without the technological process at hand.

According to Fayyad *et al.*, 'KDD[56] refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data'.[57] The necessary techno-legal conditions to execute data mining process are the following: (legal) access to the data in question, the availability of practical tools to complete the searching process and the articulation of the purpose of the process in view of an expectation.

According to existing legislation, data mining required the rightsholders' express permission because it triggers copyright and sui generis rights existing in the databases.[58] In addition, it could encroach on contract limitations imposed by the private entity that holds the data sets in question. It is often the case that the legal obstacles to getting access to the data are not confined to copyright, but that they are the result of restrictive contractual policies[59] coupled with the imposition of technological limitations and lack of interoperability or technical standards in data type formatting. Thus, the barriers that need to be overcome in order to facilitate and streamline data mining operations are not only purely legal but they are also technical and market-related. So while the market value of providing data mining services is not negligible, the existing legal framework (or absence thereof) based on private ordering and licensing formed the normative baseline that limits further opportunities.[60] Recognising the value of data mining and the fact that prima facie data mining appears to be hindered by copyright and database protection legislation, multiple examples of national laws demonstrate already implemented text and data mining exceptions to the exclusive copyrights and database rights. For example, countries such as the United Kingdom, Germany, Estonia and France have all included the exception in various forms[61] and with different requirements.[62] Recently, the text and data mining exception to copyright was adopted in the final text of the Directive on copyright and related rights in the Digital Single Market[63] voted by the European Council. According to the adopted text, 'there is widespread acknowledgment that text and data mining can in particular benefit the research community and, in so doing, support innovation'.[64] In fact,

55. The value of the effective use of data in research has been estimated in billions of euros: *See*, J. Manyika *et al., Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute, San Francisco (2011).
56. Knowledge discovery in databases.
57. U. Fayyad, G. Piatetsky-Shapiro & P. Smyth, 'The KDD Process for Extracting Useful Knowledge from Volumes of Data', 39(11) *Communications of the ACM* 27 (1996).
58. In the United States, data mining is not considered to be copyright infringement because it is qualified as fair use.
59. According to M. Dulong de Rosnay, 'right holders have been asking text and data mining to be submitted to re-licensing for an additional remuneration of texts to libraries, researchers or the public for that purpose'. *See,* M. Dulong de Rosnay, 'The Legal and Policy Framework for Scientific Data Sharing, Mining and Reuse', in C. Mabi, J.-C. Plantin & L. Monnoyer-Smith (eds.), *Ouvrir, partager, réutiliser: Regards critiques sur les données numériques*, Éditions de la Maison des sciences de l'homme (2017).
60. According to Professor Benabou, 'it is my belief that mass digitization of works – whatever the purpose is: linking, mining, crawling – implies other answers than the mere individual exclusive right and that establishing a differentiated regime of protection depending on the existence of a "sensitive" contact of the human being with a work at the end of the process would be a solution'. V.-L. Benabou, 'Text and Data Mining Issues', in *Academics Meet Policy Makers: Better Regulation for Copyright* (2017) 59, available at: https://juliareda.eu/events/better-regulation-for-copyright (last visited 15 April 2019).
61. According to Section 29A of the UK Copyright Act, making a copy of a work for text and data analysis does not infringe the copyright on the work provided that the act is made for the purpose of non-commercial research. *See*, A. Guadamuz and D. Cabell, 'Data Mining in UK Higher Education Institutions: Law and Policy', 4 *Queen Mary Intellectual Property Review* 1 (2014), at 3. According to Art. L122-5 (10°) of the French *Code de Propriété Intellectuelle*, the act of exploration of data and text associated with scientific research access to which has been obtained legally, does not encroach intellectual property rights as long as it maintains a non-commercial research goal. Also, according to Art. 60d of the German Intellectual Property Law (*Urheberrechtsgesetz*), text and data mining are permitted for scientific research absent a commercial purpose. Finally, Estonian law allows text and data mining of all types of material protected by exclusive rights, provided that the purpose of the act is not commercial.
62. While the common denominator is the use of text and data mining for non-commercial purposes, there is a divergence in the type of material covered by the exception in question. France demonstrates the most restrictive subject matter of the exception by limiting it to only text and data related to scientific research. Another divergence is also found in the requirement of prior legal access to the subject matter of the mining process. This condition is found in French law but not in the equivalent German or Estonian one. Finally, the exception in most cases covers only the right to reproduction for the purpose of the act of mining and does not include further communication to the public of the material used, or if it does, it limits its scope.
63. Art. 3 of the Directive of the European parliament and of the council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.
64. Rec 8.

Articles 3 and 4 of the EU Directive on copyright in the Digital Single Market ('DSM Directive') provide two types of exceptions and limitations to copyright for text and data mining purposes. According to Article 3, 'for reproductions and extractions made by research organisations in order to carry out text and data mining of works or other subject-matter to which they have lawful access for the purposes of scientific research'. This exception also foresees that any contractual provision preventing this operation will be unenforceable. However, the same does not apply to text and data mining activities realised pursuing commercial interests according to Article 4 of the same European text. The creation of favourable conditions towards the pursuing of research activities is evident in the European Directive.[65] Given this set of legal tools, the Commission's proposal on promoting data sharing for research purposes reflects the processes of adopting the text and data mining exceptions nationally and on an EU level.

Data – including machine-generated data – do not qualify for copyright protection because they do not fulfil the originality condition and they do not constitute human creations embodying the authors' personality. However, databases are susceptible to benefit both from copyright – if the database is deemed an original creation – and from the sui generis database right provided that there was a substantial investment made by the database owner in presenting the material of the database. Despite the absence of such rights on data, the CJEU has ruled that database owners are free to impose contractual restrictions to access on data and databases.[66] Thus, access and data mining can still prove disproportionally difficult for researchers, irrespectively of the enforcement of a text and data mining exception to copyright. Given the contractual framework that governs data use, which also reflect the asymmetries between various actors, the existing exception to copyright for text and data mining purposes is pushed to its limits. In fact, data mining restrictions are not solely dependent on exclusive rights; when state actors and public researchers inquire about getting access to privately held data sets, based on the benefits towards the public interest such as public health and environmental research, the private database holders can rely on their right to conduct a business, to claim respect for their trade secrets and to receive fair compensation. What's more, the adopted phrasing of the data mining exception links the applicability of the exception to the 'lawful access' of the researcher to the data sets in question.[67]

65. *See*, R. Ducato and A. Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to Machine Legibility', *CRIDES Working Paper Series*, 2018, available at: http://dx.doi.org/10.2139/ssrn.3278901.
66. CJEU, 15 January 2015, Ryanair Ltd c/ PR Aviation BV, aff. C-30/14. According to para. 45 of the decision, 'the Directive does not preclude the author of such a database from laying down contractual limitations on its use by third parties'.
67. From the European jurisdictions that have already implemented a text and data mining exception, the only country not imposing the 'lawfully accessed source' requirement is Germany.

Subordinating the applicability of the exception to getting legal access to a data set could significantly impact research. According to the European Copyright Society, 'the exception can effectively be denied to certain users by a right holder who refuses to grant "lawful access" to works or who grants such access on a conditional basis only'.[68]

The data set holders' strong negotiating power could lead to the inflation of the costs of granting lawful access in order to factor in the previously imposed data mining prices. In that sense, a parallel can be drawn between the mitigation of costs related to making data available for mining purposes in accordance to the Directive and the charges for the reuse of public sector documents. The PSI 2013/37/EU Directive addressed the issue of costs of making information available openly that public administrations faced. According to the Directive (and the recently published reform proposal), administrations have the right to charge for the marginal costs of making documents available and, in certain cases, they can go above the marginal costs limit if the charge is determined 'according to objective, transparent and verifiable criteria'.[69] The reform proposal adds to the following exception by determining that 'the costs of anonymization of personal data or of commercially sensitive information should also be included in the eligible cost'.[70] Similarly, and based on the public interest justification of the text and data mining exception, a framework for charges could be implemented in order to ensure preferential conditions for the effective collaboration between the private sector and publicly funded research. The fact that the private actors concerned are the sole-source data managers could contribute to the introduction of a structured and well-defined obligation for them to provide the machine-generated non-personal data under fair and non-discriminatory terms to researchers.[71]

5 Attempts at Normative Cross-Sectorial Rules of Data

The creation of access privileges to researchers is a noble goal. However, the solutions that could be implemented face challenging questions concerning the adoption of sector-specific rules or of cross-sectorial

68. European Copyright Society, 'General Opinion on the EU Copyright Reform Package', 24 January 2017, available at: https://europeancopyrightsocietydotorg.files.wordpress.com/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf (last visited 15 April 2019).
69. Rec 22, PSI 2013/37/EU.
70. Rec 32, Proposal for a Directive of the European Parliament and of the Council on the reuse of public sector information, *COM(2018) 234*, 25 April 2018.
71. Professor Hugenholtz proposed such a measure as a possible solution to overly protective contractual restrictions to databases that are not subject to copyright or *sui generis* database rights. B. Hugenholtz, 'Abuse of Database Right: Sole Source Information Banks under the EU Database Directive', in F. Lévêque and H. Shelanski (eds.), *Antitrust, Patents and Copyright: EU and US Perspectives*, Edward Elgar, Cheltenham (2005) 203.

ones, as well as the flexibility that these rules would need to incorporate taking into considerations the risks involved. Among the few attempted solutions to direct the opening of privately owned data sets, the most recent and innovative one comes from France. During the legislative process of implementing the 'Digital Republic' bill, the French legislator established normative concepts that could be further explored as an alternative solution to dealing with accessing machine-generated data.

5.1 The French Example: Public Interest Data

Mandated sharing of data exists in the form of legislation that was adopted recently in France and introduces the concept of 'public interest data'. According to the text, the objective is to 'enhance the circulation of data and knowledge' in order to give France a competitive edge in the digital economy.

The Digital Republic Bill[72] created a special category of 'public interest data' because it recognised the potential of opening up specific privately held data sets to the public for specific purposes that serve the public interest. This is the case, for example, with commercial data for the establishment of official statistics, or data relating to gas and energy consumption and production held by transmission and distribution systems operators for reuse by another party as well as data relating to changes in real estate ownership for reuse by certain third parties. In this respect, the law states that the licensor must provide the licensing authority with data using an electronic format that is open and freely reusable standard, and that 'the licensing authority or a third party designated by it may extract and freely exploit all or part of these data and databases, in particular with a view to making them available free of charge for reuse for free or for a fee'.

However, the contours of the definition of the concept remain opaque. Public interest data is not defined in the adopted legislative text, but it rather simply constitutes the title of the second section of chapter one of the Digital Republic Bill. It is therefore essential to refer to the content of the section in question – largely inspired by a report drawn up in 2015 dedicated to describing the concept of public interest data[73] – who advocated in favour of a general 'open data clause'. The Minister of Economy and Finance specified that this new concept incorporates all data 'of private nature but whose publication may be justified by their role in improving public policies'.[74] It is a significant legal innovation and it also aligns with an underlying ideological approach towards favouring access to data. The introduction of such an innovative concept is unfortunately at odds with the lack of clear definitions and guidelines as to the scope of its application. The enforceability of the provisions related to public interest data remain still largely opaque, as is the case with different data-related aspects of the Digital Republic Law.

Constituting one of the few national attempts to create a normative framework for the regulation of a data economy, the legislator takes into consideration the significant role that access to data plays in developing public policies, shaping innovation potential with respect to fairness and transparency. For this reason and recognising the need to diversify access to privately held data, the law aimed to create gateways that achieve an optimal balance between favouring market innovation and maximising societal impact. In an attempt to highlight and promote the social benefits of sharing various types of data for scientific and research purposes, the generalisation of this newly created category of data that have the potential to serve the public interest could be considered as a gateway towards better access to machine-generated data. The creation of this distinct category signals a regulatory approach towards privately held data – one that could be generalised or that could inspire a European-wide solution on the basis of the fostering of a data-based economy.

5.2 Towards a New Concept: Infrastructure Data

The public interest data definition has yet to be tested in practice in France; the ambiguity around the distinguishing elements of the concept and the scope of its application remains. For example, the scope of the public interest qualification as an autonomous concept appears to be regularly approached as a narrow definition. This is due to its nature as constituted by exceptional circumstances, applicable as an exception and not as the norm. For this reason, public interest data cannot be perceived as a general category but as an ex post qualification according to the various exceptional contributing factors. The generalisation of such a category could end up both over-burdening the concept of public interest – thus making it lose its significance – and disproportionally affect private entities that hold the data in question. The public interest nature is thus perceived as an exception to the general norm of privately held data, and, as such, it is destined to show its inefficiencies because of the elevated interest in improving access conditions to privately held data. What's more, recognising public interest as a legal justification to normalise data access can only be applied in a sector-specific way due to the diversity in machine-generated data. Thus, it cannot constitute a cross-sectorial rule.

As a way out of the dissonance between the exceptional nature of the concept of public interest and the need for exceptional access rights to diverse types of machine-generated data, the qualification of data as infrastructure

72. Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, JORF n°0235 du 8 octobre 2016.
73. Conseil d'Etat/CGE/IGF (2015), Rapport relatif aux données d'intérêt général, Inspection générale des finances – Conseil général de l'économie – membres du Conseil d'Etat, available at: https://www.economie.gouv.fr/files/files/PDF/DIG-Rapport-final2015-09.pdf (last visited 15 April 2019).
74. Ministère de l'Economie et des finances, République numérique: ouverture des données d'intérêt général, 22 September 2016, available at: www.economie.gouv.fr/republique-numerique-ouverture-donnees-d-interet-general (last visited 15 April 2019).

could create a less invasive category. The qualities of data as infrastructure have been used to justify, for example, public policies regarding Open Data. Namely, the potential social value that can be derived by accessing and reusing public data, which also possess a non-rival character, has led to the perception of open data as infrastructure provided by the public sector towards maximising social economic value and innovation.

Similarly, the potential for further innovation deriving from privately held machine-generated data has been recognised by the cases where data gathered and stored by big data companies have been used towards generating value or improving society in general.[75] The most prominent example and use case for the innovation perspectives that can be derived by granting access and use rights to the large databases of machine-generated data comes from the potential that they hold as training data sets for algorithms used in public services.[76] Within the range of the interrelations born between different data sets according to the environment in which they were gathered, the concept of infrastructure data can be developed for the benefit of the societal good in the form of scientific research. In some cases, providing access to data sets can be mandated when it is recognised that the amount of related data and the accumulated concealed knowledge potential are almost impossible to duplicate by any reasonable means for research purposes. Admittedly, the concept of infrastructure data is more easily associated with market terms in order to identify data that have a significant place in the function of a specific technology or that are imperative for further development of a technology. If in these hypotheses, fair licensing options constitute a viable solution and have been promoted as a vector for competition that results in mutual expected benefits, it has not been proven sufficient for research purposes.

6 Conclusion

This article has attempted to highlight an emerging but dominant category in the new data economy: that of machine-generated data. A growing part of current data-related literature focuses on machine-generated data from a data protection perspective. However, what this article seeks to introduce is a discussion around the implementation of rules that involve balancing of market interests, innovation, data protection and promotion of scientific advancements. In this context, the choice to focus on machine-generated non-personal data is not random. It is founded on the European Commission's proposals for the fostering of a data economy and it attempts to explore how and under what circumstances researchers can gain access to privately held machine-generated non-personal data.

Following an overview of the difficulties in delimitating the scope of application of a framework destined to apply to non-personal data, the article traces development of normative and practical approaches to the sharing of data between researchers and private entities that hold and control big databases. While we show that the applied practices have started to gain growing popularity among big companies, data sharing is far from becoming a standardised practice destined towards researchers. The need for creating legal certainty is the main impediment towards a better collaboration between research institutions and private actors. In fact, sharing data for research purposes has to ensure legal conformity with a range of property rights, private ordering clauses and the broader public good. After a description of applicable models in data sharing practices – from non-regulatory solutions (data collaboratives, data philanthropy) to regulatory ones (data mining exception to copyright, GDPR provisions for scientific research, etc.) – the article examines the recently adopted French Digital Republic Bill and the introduction of the concept of public interest with relation to data. While not undermining the potential that this concept could have should it become a more generalised category, the article underlines its shortcomings and limitations in promoting better access to machine-generated data for researchers. In fact, the qualification of public interest is a qualification that cannot be normalised without the risk of devaluing the actual concept of public interest and without risking to disproportionally affect private actors' interests. Finally, the article concludes with the concept of infrastructure data, as a similar term that could contribute towards creating data access arrangements proportional to societal needs while also taking into consideration market interests.

The goal of this article was to illustrate a range of different factors that need to be considered before attributing access privileges for research purposes. According to the needs identified, the solutions adopted have to consider whether access has to come free of charge or not and who should bear the costs of the making available. Similarly, the advantages in applying sector-specific solutions need to be assessed against the ones applying cross-sectorial ones. Before creating any type of regulatory framework that could prove to be ineffective, not flexible, and failing to respond to the needs of researchers according to rapid technological advancements, any exceptional categories created have to apply a proper balancing mechanism of interests of the actors involved towards the goal of safeguarding both market innovation and high-quality research.

75. The cases of 'corporate social responsibility' demonstrate that contribution.
76. As a matter of fact, in France, lately there has been discussion on applying the notion of public interest beyond data but also to qualify algorithms.