Some Current Aspects of Big Data Evolution

Artem A. Balyakin, Marina V. Nurbina, and Sergey B. Taranenko

NRC Kurchatov Institute, 1, ac. Kurchatov sq., Moscow 123182, Russia
Nurbina_MV@nrcki.ru

Abstract. The paper discusses the current trends in digitalization process. A number of technologies that are critical for the digital ecosystem formation (such as big data and artificial intelligence technologies) are discussed. The current approach to digital technology regulation is described. A number of characteristics of the big data life cycle are performed. The complementary connection between science and digital technologies based on it and the social life is shown. The state policy that promotes innovative development is proved to be based on setting general norms and prospects. The methods and approaches from the scientific field are supposed to be the basic tool to form digital society. A number of examples of the introduction of digital technologies in the scientific field are discussed.

Keywords: Big data · Digitalization · Megascience facilities · Ecosystem · Life cycle · Science

1 Introduction

Digitalization in modern society is considered as a classical example of breakthrough technologies. In the current century, digital technologies would play the same role as electricity did for the XXth century, enabling the authorities to solve emerging challenges. For instance, in the Russian Federation (according to the Strategy for the Development of the Information Society), the introduction of digital technologies is considered as one of the tools for solving the major challenges facing the country [1].

In addition to its obvious implementation into the technological sphere, digital technologies are also being applied in the social and economic life. The corresponding arising changes cannot be attributed to exclusively positive ones [2]: new risks and challenges contribute to the formation of digital inequality, unavoidable in the near future [3], and contradictions between existing and emerging institutions, including the degradation of social practices [4].

In this paper, the term “digital technologies” is attributed mostly to big data and data handling methods (aiming at obtaining “smart results”). We note that this approach often refers to artificial intelligence and decision-making systems that accompany big data technologies. The outcomes of this technologies development would result in a global digital infrastructure and its penetration into all spheres of human life.

Despite the variety of big data sources [5], similar approaches and algorithms are used for their processing in various fields [6]. This trait (replication property) can be treated as one of the basic characteristics of big data technologies.
2 Big Data Concept

From a legal point of view, “big data” is not a legal term at the moment, but rather a collection of approaches with many different implications for various disciplines such as economics, natural science, legal and social sciences, etc. [7].

The Russian Federation emphasizes the position that the main source of big data is the Internet and therefore “big data” are defined as Big User Data [8]. According to the conclusion of the US Federal Trade Commission, “big data” are arrays of structured or unstructured data characterized by large volume, variety, high rates of change, and real-time processing [9]. In European Union main attention is paid to socio-economic aspects of big data turnover, and no specific legal definition has been accepted. In fact, big data can be defined as a set of data and information that defies ordering and sorting at the present stage of human development.

This approach can lead to legal uncertainty in this area. In our opinion, the most applicable definition should emphasize a new qualitative property of large amounts of information (rather than quantitative ones), considered as “big data”, in comparison with the usual datasets. Thus, big data are perceived as a complex that includes both the data itself and the approaches and methods of their analysis. In this case, big data are rather a process, not an object.

3 Perception of Digital Technologies

The abovementioned shows the high role of digital technologies in general (and, more narrowly, big data technologies) for the social sphere [10]. The growth of technical capabilities greatly impacts the social space. Contrarily, at the same time, the changes caused by innovative technologies are institutionally dependent [4, 7]: their possible negative and positive consequences are tied to the proper and timely development of social institutions.

In numerous works related to the field of big data, technical issues of organizing the collection, storage and processing of information prevail. The socio-economic and humanitarian aspects of high technologies are presented much worse. For example, the author [11] gives big data technology a “humanitarian tint”, proposing to use the concept of cultural and social capital, developed by P. Bourdieu, in relation to big data. In his opinion, big data, like cultural and social capital, simplifies interaction with the information environment, which creates conditions for the conversion of capital into economic goods, the last to generate net income.

General philosophical questions of the algorithms creators responsibility are discussed, for example, in [12, 13]. The authors proceed from the fact that algorithms silently structure our life, removing responsibility from officials, “depersonalizing” decisions, believing it to be a common good [10]. Instead of reliability and objectivity to other uncertain procedures, the sensitivity of algorithmic models to human bias increases [13]. The solution is seen as a deeper empirical study of algorithmic models used in practice [13], and the growth of “internal responsibility”, when it is proposed to increase the involvement of employees who develop algorithms, making them ethically responsible for the solutions developed [12]. This approach looks very controversial, since it does
not mention the responsibility of the “customer” who sets technical tasks. As a rule, it is the state [14], which, depending on its goals and objectives, implements different approaches (and, as a result, demonstrates the demand for different algorithms). For example, building a digital economy based on blockchain technology in Estonia is more successful and meets less resistance in society than in the UK, where cultural traditions, coupled with a larger population, create problems for building the so-called “algorithmic government” [14].

As a rule, many studies touch on “borderline” issues related to improving the technical characteristics of systems, leading to its qualitative growth. In particular, the authors of [15, 16] showed that an increase in the number of data used does not mean an improvement in the output results. The problem is associated with both nonlinear saturation [16] and an increase in the influence of noise with increasing data set [15]. On the other hand, the algorithm used in big data technologies itself generates “new” big data [11]: the collection of more and more data makes the information environment more and more controlled and manageable, which in turn creates the opportunity to collect even more data. This trend does not lead to qualitative changes yet, and the redundant data is actually not of interest. They are sometimes necessary however for the formation of more fine-tuned decision-making mechanisms (for example, in the field of sales and service delivery).

The solution to emerging difficulties is seen both in new algorithms and approaches [15, 16], and in the development of new ethical rules and norms related not only to the developers of algorithms (as suggested in [13]), but also to the system as a whole. Thus, in [17], using the example of the work of international scientific collaborations (including megascience-class facilities), it is proposed to use the rules for protecting the emerging intellectual property of project employees as a starting point, which are previously developed and approved by the project participants.

The authors stress the complementary relationship between science and digital technologies (rooted from the scientific approaches and tasks) and the social life. There is a bilateral influence of one on the other [10].

On the one hand, the algorithms used are a priori biased, and on the other hand, the results obtained seriously change the socio-cultural landscape (therefore, provoking changes in algorithms in return). For example, in the legal field, there is a deformation of law due to the delegation of decision-making from a person to algorithms [4]. In a practical aspect, this is realized by training algorithms on existing practices (precedents). As a result, the emerging artificial intelligence (decision-making device) can accumulate both the norm and negative practices. This can be avoided only by setting rules of a higher level than the algorithms themselves [5, 17].

4 Evolution of Big Data Technologies

4.1 From Big Data to Smart Content

Understanding digitalization phenomenon requires studying the evolution of technologies involved in this process. So, in relation to big data and artificial intelligence, there is a process of “data intellectualization”, which has received the name of the phenomenon of “smart content” (vs “habitual” big data). It has the following features:
1. Knowledge vs new data. The task of such a process is to get away from a simple search for correlations, which is typical of most methods of working with big data. The results obtained should be categorized as “knowledge” (as a result of data interpretation) and not “new data”.

2. Knowledge vs code. Data is thought of as an “active participant” in the processing and decision-making process. The idea was expressed to combine “data” and “code”. In particular, in accordance with the FAIR principle [18], algorithms are included in the “circle” of stakeholders and participate in decision-making (as an independent object, whose “vote” - the decision obtained - is counted with a certain weighting factor).

3. An priori assignment of good and evil. The social component of big data technologies is taken into account: solutions based on digital technologies are not socially neutral, and must be “adjusted” in advance in the direction necessary for society. Accordingly, the contradiction between personal data and open data could be removed. On the other hand, it opens wide opportunities for fake results and eliminates the unwanted results (thus inclining into socially committed results). To manage that process the human control proves to remain inevitable.

4. Quality vs quantity. On the “technical” side, smart data will be characterized by fewer variables than currently, along with the ability to fold and recover data: the amount of data will decrease without loss of quality. This thrusts the need of new algorithms development, and smart data storage.

In our opinion, smart content is the data of the subject area based on big data, on which, despite their immensity, an adequate interpretation is built due to immanent intelligent procedures. From the user’s point of view, smart data is in some way an intermediary between digital technology and civil society (an understandable representation of data set).

4.2 Big Data Life Cycle

The life cycle as an analytical concept is characterized by life time, and understood as a characteristic time of significant changes. Hereinafter we list a number of the big data life cycle features in the aspect of legal regulation.

- Taking into account the dynamics of the context (not that all is big data that seems to be). The very concept of “big data” is dynamic. The current development of deep learning technology, and then the expected development of universal AI, as well as the possible development of cognitive (nature-like) AI, will significantly change what we can design (create, process, use) as big data. At the same time, some of what we previously understood as big data will lose this status, becoming a traditional database.

- Accelerating social change. Big data is conceived according to current trends (fashion-like style), and as social life is tremendously unstable and ever altering, so does the big data differs. In practical aspect, we mention the requirements regarding the “right to be forgotten”, etc., that may change dramatically within a short period of time.
Integration of big data. Big data has both an opportunity and a tendency to be part of larger sets (a kind of fractal structure). The latter is clearly manifested, first of all, in retail and banking.

Multiplication of characteristic life cycle times (presence of several characteristic times for one phenomenon).

Dynamism of definitions and legal mechanisms during the life cycle. According to the authors, the resilience of big data to these changes, including the need to support them (functioning, updating, legal adaptation, etc.) should be included in the analysis of the big data life cycle.

Taking into account consideration of the upstream and downstream stages of the big data life cycle. In the early stages, big data may not meet the big data requirements. Moreover, at this stage they are collected and processed. An independent legal regime is needed. In the final stage (refusal of support, modification, etc.), not only atomized rights and needs (necessities) may be affected, but the functioning of systems and infrastructure may also be disrupted. With the possible absence of an alternative (both physically and economically possible), this is a matter of national security.

To our viewpoint, the big data life cycle is a set of stages, various forms and pace differenced processes in the pregenesis (initial stage, as a creation of the idea of big data and corresponding issues), genesis and post-genesis (functioning of big data) as a systemically connected and organized institution. By now the big data technology has not reached the stage of mature technology, and came to a stand at the position of the tool (for other scientific and industrial areas). Its further evolution is still under question, and the course has not been settled yet.

5 Conclusion

Big data technologies have already outgrown the role of a tool for solving a narrow circle of strictly applied problems to an almost independent scientific and technological field, which in the minds of most people is rigidly linked with artificial intelligence and decision-making systems. This complex phenomenon is better treated as a process, rather than an object. Thus, being a process, it gives rise to a number of risks and challenges. To overcome them a number of actions is required.

Adequate legal regulation in the digital sphere should be one of the first steps [19]. Understanding the essence of the processes caused by digitalization leads to the idea that it is society (represented by the state) that sets the rules and regulations for business, and not vice versa. To preserve social well-being and development, it is necessary to ensure the creation and development of a well-thought-out and effective regulation system through the law of relationships between man, science, state, society and nature [20].

It is necessary to understand the dynamics of development of digital technologies, taking into account the peculiarities of the life cycle of big data and other technologies in strategic planning. This leads to the need to separate legal mechanisms at different stages of building the digital economy.

For example, in the scientific field, digitalization processes are clearly represented in the work of megascience class facilities [21]. Particularly, the coronavirus pandemic has
given an additional impetus to the widespread introduction of the remote access regime and the creation of digital twins of real objects [20]. This phenomenon is still under development.

According to the authors, there is no need for immediate legal regulation of all emerging technologies at the time of their creation. The main task of states interested in the development of advanced digital technologies is to formulate meanings and set a perspective. An example is national strategies for the development of artificial intelligence, digital technologies, etc. Perhaps part of such a regulatory system belongs to the newly formed complex branch of law - scientific research law (the law of science) [22].

In this regard, the authors actively support the idea of developing a set of supranational/international principles of human interaction in connection with the creation of AI systems, including the ethics of the application of advanced digital technologies and the separation of the law of science into a separate industry. We believe that such a set of rules can be based on the Report on the Ethics of AI (2019) of the World Commission on the Ethics of Scientific Knowledge and Technology of UNESCO [23].

Acknowledgments. This work was supported by the RFBR grant 18-29-16130.

References

1. Decree of the President of the Russian Federation of 09.05.2017 No. 203. On the Strategy for the Development of the Information Society in the Russian Federation for 2017–2030. https://pravo.gov.ru/proxy/ips/?docbody=&firstDoc=1&lastDoc=1&nd=102431687. Accessed 20 Oct 2020
2. Insight Report. The Global Information. Technology Report 2016. Innovating in the Digital Economy. https://www3.weforum.org/docs/GITR2016/WEF_GITR_Full_Report.pdf. Accessed 01 Nov 2020
3. Zobova, L., Shecherbakova, L., Evdokimova, E.: Digital spatial competition in the global information space. Fundam. Res. 5, 64–68 (2018). [in Russian]
4. Zhulego, V.G., Balyakin, A.A., Nurbin, M.V., Taranenko, S.B.: Digitalization of society: new challenges in the social sphere. Bull. Altai Acad. Econ. Law 9–2, 36–43 (2019). [in Russian]
5. Balyakin, A.A., Malyshev, A.S., Nurbin, M.V., Titov, M.A.: Big Data: Nil Novo Sub Luna. In: Antipova T. (ed.) Integrated Science in Digital Age. ICIS 2019. Lecture Notes in Networks and Systems, vol. 78, pp. 364–373. Springer, Cham (2020)
6. Grigorieva, M., Golosova, M., Ryabinink, E., Klimentov, A.: Exabyte repository of scientific data. In: Open Data Systems. DBMS, Moscow (2015). https://www.osp.ru/os/2015/04/13047963
7. Perspectives in Law, Business and Innovation, p. 341. Springer Nature Singapore Pte. Ltd. (2017)
8. National Program “Digital Economy of the Russian Federation”. https://government.ru/info/35568/. Accessed 02 Sept 2020
9. Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues (FTC Report). https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report. Accessed 04 Oct 2020
10. Balyakin, A.A., Taranenko, S.B., Nurbin, M.V., Titov, M.A.: Social aspects of big data technology implementation. ICS. J. Digit. Sci. 1(1), 15–24 (2019)
11. Sadowski, J.: When data is capital: datafication, accumulation, and extraction. Big Data Soc. 6(1), 1–12 (2019)
12. Kirsten, M.: Ethical implications and accountability of algorithms. J. Bus. Ethics 160(4), 835–850 (2019)
13. Kemper, J., Kolkman, K.: Transparent to whom? No algorithmic accountability without a critical audience. Inf. Commun. Soc. 22(14), 2081–2096 (2019)
14. Engin, Z., Treleaven, P.: Algorithmic government: automating public services and supporting civil servants in using data science technologies. Comput. J. 62(3), 448–460 (2019)
15. Triguero, I., García-Gil, D., Maillol, J., Luengo, J., García, S., Herrera, F.: Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. WIREs Data Min. Knowl. Discov. 9(2), e1289 (2018)
16. Succi, S., Coveney, P.V.: 2019 Big data: the end of the scientific method? Phil. Trans. R. Soc. 377(2142), 20180145 (2019)
17. Kahn, M.: Co-authorship as a proxy for collaboration: a cautionary tale. Sci. Public Policy 45(1), 117–123 (2017)
18. The FAIR Data Principles. https://www.force11.org/group/fairgroup/fairprinciples. Accessed 03 Aug 2020
19. Balyakin, A.A., Nurbina, M.V., Taranenko, S.B.: Comparative legal features of the formation of a digital ecosystem. Int. Legal Courier 1–2(37–38), 42–45 (2020)
20. Nurakhov, N.: The basic processes of creating a "Megascience” project. In: International Conference on Integrated Science, ICIS 2019: Integrated Science in Digital Age, Batumi, Georgia, pp. 329–339 (2019)
21. Nurbina, M.V., Nurakhov, N.N., Balyakin, A.A., Tsvetus, N.Yu.: Mega science projects for business. In: Aham, T., et al. (eds.) Human Interaction, Emerging Technologies and Future Applications III Proceedings of the 3rd International Conference on Human Interaction and Emerging Technologies: Future Applications (IHIET 2020), Paris, France, pp. 488–492 (2020)
22. Kashkin, S.Yu.: The formation of the law of science as a new integrated branch of law, Bull. Univ. Named After O.E. Kutafina, Moscow (5), 16–27 (2018)
23. Report of COMEST on robotics ethics. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000253952. Accessed 09 Nov 2020