Big data technologies in environmental monitoring

Svetlana Yu Piskorskaya\textsuperscript{1} and Yulia N Malanina\textsuperscript{2}

\textsuperscript{1} Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russia
\textsuperscript{2} Irkutsk State Transport University, 15 Chernyshevsky St., 664074, Irkutsk, Russia

E-mail: yuliamalanina@mail.ru

Abstract. The article deals with the big data technologies currently used for environmental monitoring. The processes of accumulation and preservation of super-large data sets, technologies for working with them, and a paradigm shift in science are presented. A predictive function of big data has been identified that allows effective use of this technology in environmental monitoring. A general idea of environmental monitoring as a complex system of observations, assessment and forecast of changes in the environmental state has been formed. The main areas of environmental monitoring are identified. As an example, the activity of collecting, processing and storing big data in the framework of the strategic project "Geographic Information system "Yenisei-Arctic", implemented by the employees of Reshetnev Siberian State University of Science and Technology. The main tasks and results of the project are presented. Among the main results of the project, the formation of a research direction in the field of big data and the involvement of young scientists and students in promising research that can prepare the necessary information for the needs of environmental monitoring is highlighted. It is proved that big data technologies can raise environmental monitoring to a whole new level.

1. Introduction

In the modern world, the amount of statistical information describing natural processes and phenomena that are heterogeneous in composition and characteristics, as well as the availability of methods for obtaining them, increases every year.

So, if in 2000 the amount of information stored in digital format was measured in exabytes and amounted to twenty-five percent of the total amount of information in the world, in 2007 it is measured in zettabytes (a thousand times more), ninety-three percent of the total amount of which was represented in numbers. In 2013, the amount of stored information in the world was 1.2 zettabytes, of which the digital format accounted for more than ninety-eight percent [1]. This dynamic in the amount of stored data clearly demonstrates that the digital world is rapidly expanding.

How much data is considered large (big) is determined by the technical capabilities. The term Big Data itself is already widely used in modern digital society. We voluntarily surround ourselves with digital networks of various kinds of devices recording information. Such devices can include different sensors, cameras, scanners, satellites generating unstructured data arrays. The widespread use of Big Data is irreversible because it is a requirement of time, new technological order and technological cycle.

First proposed by John R. Mashey as chief scientist at Silicon Graphics in 1998, the term Big Data was not adopted by the scientific or professional community. In general meaning, by the term Big Data, we consider a set of data with possible exponential growth. These data are too volume, big, are not
formatted and not structured for the analysis by traditional methods and means. The widespread use of analytics provides new and broad perspectives for the creation of artificial intelligence technologies for new approaches to economic management, agriculture, mechanical engineering, metrology, etc. However, for all this to become possible, simply the development of technologies, as it was before, is not enough. New development will require specialists with not only mathematical and engineering knowledge, but also systemic thinking.

With the new century, the situation has changed fundamentally. All modern computers have been connected to the network, and therefore to an unlimited amount of the most diverse content. This content consists of "old" structured data and "new" unstructured data. Similar to the past, classical methods of working with data could now be called "package." This data, as it used to be packages, is downloaded to the computer and then processed. Such approaches remain relevant. However, a variety of analytical technologies are becoming significant allowing to extract useful information from large amounts of data in online time. The second data source is related to the so-called “sensory revolution”, which brought with it streaming data and the need to process it (streaming processing) and highlight significant events (complex event processing). Today, any digital device is equipped with sensors capable of transmitting data over the network.

In fact, big data is intended for forecasting by applying mathematical methods to a large amount of data to calculate the probability of an event occurring. These systems are designed to improve their forecasts over time by providing new data about the objects, phenomena, or processes being studied. As a result, there is a transition from micro-level thinking to macro-level thinking - from the requirement of accuracy, concreteness and causality required at the micro-level, to global assessment, finding a correlation between data that allows you to make fundamentally new discoveries at the macro level.

2. Methods and results

Note that identifying patterns in large amounts of data becomes one of the main methods for obtaining new knowledge. When creating programs and applications for working with big data, developers face a number of difficulties. These difficulties are related to such factors as: extremely large data amounts, structural complexity and heterogeneity of data, increasing weakly structured and fuzzy source information, increasing demand for parallel computing, the requirement to increase the speed of data processing and analysis, and the limit of decision-making time for any number of data.

Programs focused on processing large amounts of information work with files ranging from a few terabytes to a petabyte, and the data is presented in different formats and on different sources of information storage. Their processing appears as a kind of step-by-step pipeline: from transformation to integration and subsequent data analysis. At the same time, the constant increment of information requires an increase in the speed of its processing [2].

Big data technologies help you quantify the world around you. And what was previously impossible to comprehend because of the large volume, now you can not only measure, but also store, process, and distribute. These technologies will help solve the main tasks of environmental monitoring to solve the global environmental problem (for example, in terms of climate change) [3].

This is due to the fact that environmental monitoring is a comprehensive system of observations of the state of the environment, assessment and forecast of changes in the state of the environment under the influence of natural and anthropogenic factors, as well as informing authorities and the population about the state of the environment.

Activities in the field of environmental monitoring include obtaining measurement information about the state of controlled natural environments; evaluating the environmental state of natural environments, analyzing the current environmental situation and forecasting its development [4].

The environmental monitoring system involves the analysis of large unstructured data, which includes tables, videos, photos, text documents, and graphs. it allows you to detect patterns that are difficult to catch in the classical approach. This is achieved by establishing a correlation between the collected data.

However, with all the advantages of big data technology, it is necessary to understand the risks
associated with the possibility of false conclusions, since mathematical models embedded in algorithms can detect not a pattern, but only a partial dependence and make an incorrect conclusion, but this does not contradict the idea of probabilistic knowledge obtained on their basis.

This also proves that the use of big data technologies allows you to raise environmental monitoring to a new level, in a new way shaping its content, properties, parameters and structures of accumulated information, technologies for its processing and data management, multidimensional modeling and presentation on the Internet. The resulting information space becomes part of the general information field and, based on big data analysis, allows you to display the properties of the surrounding world [5].

Let's consider such activities on the example of the strategic project "Geographic Information System" Yenisei-Arctic", implemented by the employees of Reshetnev Siberian State University of Science and Technology. The goal of this project was to create a geographical information system for the lower reaches of the Yenisei river, its inner delta, and the Yenisei bay. The system consists of maps layered on top of each other, which allows the user to get a variety of information about the hydrography of the Yenisei. In the future, it is planned to collect additional information that is valuable for analyzing the environmental situation of water areas.

It should be noted that the idea of creating this project appeared in the team of teachers-historians of the University. From 2015 to 2017, they implemented research projects related to the history of navigation on the Yenisei. Two seasons of field research were conducted, during which it was possible to localize four sunken ships on the Lower Yenisei associated with the history of the development of the Northern sea route. Among the most significant finds were the English steamer Phoenix (sunk in 1892) and the English steam schooner Thames (sunk in 1878). These objects were found after analyzing historical sources, which allowed us to establish the location where these sea ships were last seen.

The research was carried out using a side-view sonar device for detecting underwater objects. To date, acoustic instruments are the only instruments that can obtain satisfactory data about the bottom of the reservoir [6]. The side-view sonar sends acoustic signals at a specified frequency and processes feedback signals that are reflected from underwater objects and the bottom of the reservoir. The processed data is sent to the user's screen as a graphical image. With proper operation of the device and suitable weather conditions, the image quality can be very good and allows you to see the smallest objects.

This made it possible to approach the study of sunken ships on the Yenisei river from a broader perspective, studying not only any ships and their history, but also the state of the waters of the former capital and non-capital ports, as well as shooting large areas of the Yenisei river in order to monitor its environmental status. The storage and processing of such a volume of information required large computing power.

The solution of this issue formed the basis of the University project "Geographical Information System "Yenisei-Arctic". The research team includes specialists in the field of history, specialists in geographical information systems, geodesy and remote sensing of the earth, programmers and specialists in big data, applied mathematicians, specialists in the use of water resources, specialists in the field of robotics, specialists in applied computer science, and others. This made it possible to consolidate the efforts of specialists on various scientific topics to solve one applied problem – the construction of a geographical information system. The main objectives of the project were:

- collecting and processing of a large amount of data (satellite images, acoustic images of the bottom, bathymetric data);
- collecting and processing of historical data (old maps and lots (up to 100 years old), archival data, historical photos, etc.);
- building models of the geographical information system in the form of separate layers and mosaics in the corresponding computer programs;
- development of an uninhabited remote-controlled underwater vehicle for spot study of underwater objects;
- creation of a unified database of sunken man-made objects;
• creation of a database of underwater and land historical objects related to the development of the Northern Sea route and the development of navigation on Yenisei;
• development of methods for research and exploitation of sunken wood on Yenisei;
• determination of the dynamics of geomorphological changes in the channels of the Yenisei delta on the basis of dynamic models of the geographical information system. This method is the author's and was described earlier [7].

In addition to these tasks, during the project implementation, it was decided to develop special mathematical models based on the Bayesian search theory and other statistical methods for more efficient search of underwater objects. This idea is then developed in a solution to develop specialized software to optimize search any wrecks that will be very useful for search and rescue services such as emergency, operating in the North. It should be mentioned that such software has been used for almost two decades by American services to search for sunken small vessels [8].

The implementation of the main part of these areas is directly related to the area of big data. So, during the field work of 2018, data was collected in the amount of about 30 terabytes. At the same time, it is necessary to pay attention to the fact that the use of foreign software for processing geodata is a security concern. For example, foreign manufacturers of affordable hydroacoustic devices, such as Garmin, Lowrance, Humminbird, and others offer to use corporate software for processing collected data and building bathymetric maps. Some services require sending data to a remote server, where it is sent to the user in the form of ready-made maps. What happens next with files stored on the server is unknown. Files with the extension DXF, SHP, KML, CSV, GPX, SON, and others are private and cannot be decrypted by standard programs [9].

In such circumstances, it is necessary not only to develop national domestic scientific instrumentation in the field of hydroacoustics and underwater robotics, but also to develop specialized software for these scientific instruments. The Yenisei-Arctic Geographical Information System project is aimed at partially eliminating these problems by manufacturing its own underwater vehicle [10].

Today, among the main results of the project, we can single out the formation of a research direction in the field of big data and underwater robotics; attracting young scientists and students to promising research, capable of preparing the necessary information for the needs of environmental monitoring.

That is why this project can serve as one example of information systems during environmental monitoring, since monitoring, auditing, and the introduction of environmental innovations, as well as the development and implementation of environmental programs and projects, require the storage and processing of large amounts of information. After all, environmental monitoring is essentially a multi-purpose information system of long-term observations, as well as assessment and forecast of the state of the natural environment. The main purpose of environmental monitoring is to prevent critical situations, harmful or dangerous to human health, well-being of other living beings, their communities, natural and human-created objects. The monitoring system itself is a source of information necessary for environmental decisions.

The same applies to environmental forecasting by applying mathematical methods to a large amount of data to calculate the likelihood of events associated with various changes in human impacts on the environment. Mathematical methods have integrated a wide variety of fields of theoretical and applied ecology and ecological forecasting: in the analysis of species relationships in the community, in the study of migration processes, territorial behavior, in the analysis of substance and energy flows in ecosystems, in the study of problems of complexity and sustainability of communities, as well as assessments of the impact of various anthropogenic factors on natural systems, in the study of problems of optimal management of natural resources and exploitation of populations, etc. In the current conditions of intense anthropogenic impact on the natural environment due to the high level of technological development and more active interference of human activities in natural processes, it has become necessary to establish a system of observation of the state of the environment in order to monitor, predict and manage this state. The data (Big Data) obtained should form the basis for the development of an environmental impact assessment of the planned activity and the development of a
system of environmental measures ensuring the permissible degree of impact and, if necessary, measures to improve the environmental situation in the planned territory.

3. Conclusion
The modern solution of local and global environmental problems requires new approaches, including the involvement of big data technologies in the practice of environmental monitoring. Big data includes three components: the data itself, the technologies for working with them, and the new paradigm in science, which can be successfully applied in an integrated system for observing the state of the environment, including the assessment and forecast of its changes.

References
[1] Maier-Shenberger V 2014 Big Data Revolution That Will Change the Way We Live, Work, And Think (Moscow: Mann, Ivanov, and Ferber) p 240
[2] Chekharin E E 2016 Big Data – big problems Science and Education Prospects 3(21) 7–11
[3] Big data for governmental management 2017 Available at: http://ac.gov.ru/files/content/10087/pushkin-bolshie-dannye-dlya-gosupravleniya-pdf.pdf
[4] Piskorskaya S Y and Malanina Y N 2019 Environmental engineering: regulation, monitoring, audit, innovation Journal of Physics: Conference Series 1399 033057
[5] Katsko S Yu 2014 Big Data for geographical information system spaces for effective management at situations of crisis Available at: https://cyberleninka.ru/article/n/bolshie-dannye-v-geoinformatsionnom-prostranstve-dlya-efektivnogo-upravleniya-v-krizisnyh-situatsiyah
[6] Blondel P 2009 The Handbook of Side Scan Sonar (UK: Spinger–Praxis Publ. Ltd) p 316
[7] Van Der Wal D and Pye K 2003 The use of historical bathymetric charts in a GIS to assess morphological change in estuaries The Geographical Journal 169(1) 21–31
[8] Malakoff D 1999 Bayes offers a ‘new’ way to make sense of numbers Science 286(5444) 1460–4
[9] Badmaev A B, Kozienko L V and Klimov N N 2019 On the possibility of using single-frequency GPS/GLONASS receivers during shunting operations at the station Modern Technologies. System Analysis. Modeling 2(62) 212–20
[10] Piskorskaya S Yu and Goncharov A E 2018 Big Data as a phenomenon of social philosophy and professional education Professional Education in the Modern World 8(4) 2178–85