Methodology and practice of development of Data Analytics Competencies

Olga A. Fiofanova
Russian Academy of National Economy and Public Administration under the President of the Russian Federation, Moscow, Russian Federation, fiofanova-oa@ranepa.ru

Mikhail S. Averkov
Novosibirsk State Technical University, Novosibirsk, Russian Federation, averkov_m@yandex.ru

Alexander A. Popov
Russian Academy of National Economy and Public Administration under the President of the Russian Federation, Moscow, Russian Federation, popov-a@mail.ru

Pavel P. Glukhov
Russian Academy of National Economy and Public Administration under the President of the Russian Federation, Moscow, Russian Federation, glukhov-pp@ranepa.ru

ABSTRACT

In the context of the digitalization and the implementation of big data technology, the demand for the development of data-analytics competencies, competencies in the analysis of digital data is becoming actual. The relevance of the article is in identifying new types of educational practices - practices of working with data. Purpose of the article: on the basis of a methodological analysis of concepts about data, to identify and systematize conceptual approaches and educational practices of working with data. Research methods: methodological analysis, method classification, structural and functional analysis of concepts and practices for the development of data competencies. In the process of methodological and structural-functional analysis, educational practices for working with data were identified and systematized: a) practice that ensures the implementation of a research project based on working with data; b) practice that have potential to master working with data as a complex learning activity;c) practice ensuring the use of data analysis as a means of working with other types of knowledge, with other types of academic subjects; d) practice of working with personal educational data, ensuring self-determination and educational individualization of students. The research results contribute to the development of data science in the methodology of pedagogy and education, in the pedagogy of competence development.

KEYWORDS

Data science, big-data in education, data-competence, educational practices for the development of data-competences

1 Introduction

Over the past decade of development of education, methods of data science are increasingly moving into educational practice, especially intensively in the conditions of digital transformation of education.
The methodology for analyzing large data in education is used in two contexts:

1. in the context of the application of the Data-Driven approach in education - analysis of data (educational results) to organize education based on data and evidentia education;
2. in the context of constructing the content of education as data analysis practices when studying certain topics, research and study projects of students.

It develops date-literacy among students, increasing their competitiveness in the world of digital economy.
From the teacher, this requires new competencies - date-competences. Data competence of teachers is implemented both in the aspect of the organization of education on the basis of the analysis of educational data and in the aspect of the didactics of the detention of education - the use of data on the development of various sectors of the economy as a content of educational topics or the content of design topics in educational activities.
The method of data analysis, including in large arrays, has long been the subject of development within the education system, including the subject of educational order from parents, students, other structures interested in the competence development of students. Such a situation is fully explained, since, first, one of the most vivid characteristics of the cash cultural situation is the increasing amount of information, which should operate a specific person; Secondly, digital competencies developed not only on everyday life, but also at the production level, they determine the high competitiveness of the young professional. But at the same time, in our opinion, the possibilities of working with large data in the context of modern education are not used sufficiently.

The expansion of the information sphere makes the student's understanding of traditional knowledge based on any source: a textbook or a teacher insufficient. The open source information retrieval approach is not enought because it does not assume that the seeking student has sufficient theoretical or practical goals for the search, as well as an idea of the structure within which the material will be built. These shortcomings can be compensated for through the use of data analysis methods - but only if it ceases to be interpreted as a section of objective knowledge and begins to position itself for students and offer them as a methodology of cognitive activity.

2 Materials and methods

The study used methodology developed in the framework of the project on a grant of the Russian Foundation for Fundamental Research - «Methodology for the analysis of bulk data in education and its integration into training programs for teachers and heads of educational institutions in the logic “Pedagogy based on data”, “Management of education based on data”» [1]. Research methods: Methodological analysis, method of factor analysis, systematization and classification method, structurally functional analysis of the development of data-competences. The results of the study contribute to the science of data into the methodology of pedagogy and education, in the theory of general and vocational education.
The study used the following sources of materials:

1. Publications in scientific and methodological journals describing specific precedents for the use of working with data and data science in the educational process - as its direct methodological toolkit, as sources of information for making pedagogical or managerial decisions, as tools for the implementation of individual (tutor) accompaniment.
2. Digital resources dedicated to specific educational practices that use data manipulation as a significant tool for their implementation.
3. Digital resources dedicated to specific developments in the field of working with data and their capabilities, including in the field of education.

The research is based on the methodological principles of cultural-historical, activity-based, systemic approaches in the humanities, suggesting that educational effects, including those associated with the development or use of complex tools, can be provided only by a specially organized multi-position and multifactor space.
During the research the following methods were used: systemic genetic analysis of the phenomenon (including the didactic principle or method); system modeling of the object under study, including the reconstruction of its implicit components with subsequent verification; comparative analysis of the studied phenomena.
Research questions:

- what are the educational practices that ensure the acquisition of data-competences?
- what are the educational practices that ensure the use of data analysis as a means of working with other types of knowledge?
- what are the educational practices that ensure self-determination and educational individualization of students?

3 Results

Key in the research is the concept of "Data literacy", which is expressed in the ability to understand the essence of data and use it as a tool for solving cognitive and practical problems, as well as the ability to critically treat information through a hypothetical reconstruction of the datasets underlying it. To define this concept and its meaningful connotations, we relied on the works of M. Schield [2], J. R. Carlson [3], E. B. Mandinach and E. S. Gummer [4]. At the same time, some authors, for example A.F. Wise (Wise A.F., 2020) views the realm of data science and data literacy as a critical theory. This approach is based on the "productivity" mindset, and it is assumed not so much the deepening of technical skills in data analysis as the massive development of "critical data literacy" to assess the existing social situation and form a personal civic position.
When defining the main models of using artificial intelligence in educational activities, we relied on the works of R. Luckin, W. Holmes, M. Griffiths, L.B. Forcier [6]. The authors build their typology on the basis of those basic objects that are represented by means of artificial intelligence: pedagogical design, understood as a system of pedagogical principles, technologies, criteria and forms of educational expertise ("pedagogical" model of using artificial intelligence in education); subject area, reconstructed and mastered with the help of artificial intelligence ("subject area model"); the totality of the student's personal characteristics, including his educational preferences, basic attitudes, abilities, on the basis of which competencies can be formalized ("student model"). The existence and use of each of these models of using artificial intelligence in education knowingly forms the corresponding didactic models, which can be conventionally called "spatial-event", "subject-content", "personal-anthropological". But all of them exist in the mode of a large number of private precedents, the descriptions of which do not go beyond fixing specific methods of work, so this key aspect of including work with data in pedagogical practice remains practically not covered in the special scientific and methodological literature. At the same time, the issue of key types of data is considered in detail, work with which will be most productive for achieving a particular educational result, in particular, for the formation of data literacy, accompanying metasubject competencies (in particular, critical thinking), reconstruction of the structures of subject knowledge. One group of specialists believes that the greatest effect here will be obtained by working with authentic scientific data (Kjelvik M. K., Schultheis E. H. [7]). On the other hand, the authors of the widely cited approach of "data steps", in particular, T. Erickson [8]. believe that an introductory data science course should provide students with problems with varied content, raw, “naughty” data. At the

same time, since “complex” data is not necessarily the same as “big”, and the concept of “real data” does not necessarily mean “authentic data”, the concept of “authentic data”, considered in detail by Kjelvik and Schultheis, seems to be productive.

Also, in our study, we based on scientific works devoted to the role of working with data in the formation of certain types of thinking, that is, in fact, complexes of meta-subject competences. A. Cuoco, E.P. Goldenberg, J. Mark [9] substantiate that the “mental habit” of accessing data is knowingly transformed into an appropriate way of thinking, based on the construction of information models and on the operation of them. W. Finzer [10] actually goes further and discusses the formation on the basis of working with data not just a sustainable approach to solving problems, but a certain picture of the world - “looking at the world through the prism of data”.

The use of working with data to build an educational environment and manage the educational process was analyzed in detail by J. Wong et al. The comparative table compiled by them makes it possible to determine what types of educational and educational spaces today actually use data analysis to ensure the most effective management, what data have to be worked with in each case, what analysis methods are used for this. J. Wong and his co-authors come to the conclusion that working with data can become a management resource for those educational practices that rely on the theory of self-regulatory learning, on specially organized work with motivation based on various approaches, on the approaches of social constructivism. This also includes works on data mining in education (Educational Data Mining, EDM). It provides extraction of previously unknown non-trivial and practically useful knowledge from educational statistics, revealing hidden patterns that allow drawing conclusions and making predictions about student performance.

The analysis of scientific works allowed us to distinguish two groups of developments related to the study of the problems of using big data in education. The first group explores the application of big data analysis methods in an organizational and managerial context in order to optimize the educational process, support personal learning trajectories and improve academic performance. The second group considers data in a didactic context, as an object, means or content that learners master.

In the course of the study, we identified and analyzed both a significant array of complex educational practices devoted to data analysis, and a significant number of private methods and techniques for working with data that are successfully used to organize the educational process. They can be conditionally divided into the following groups, in accordance with the objective function:

a) practices that ensure the development of competence in working with data;
b) practices that have potential to master working with data as a complex activity;
c) practices that ensure the use of data analysis as a means of working with other types of knowledge;
d) practices that ensure self-determination and educational individualization of students.

At the same time, in the study, we deliberately did not single out as a separate category of practice in which data analysis would become a resource for competence development, since this function is actually performed by any practice that provides work with data in an activity approach. Similarly, we did not single out the practice related to the use of data for managing the educational process and for reconstructing the student's personal situation in a special section.

Practices in mastering data skills.

In many cases, data curricula targeted at adolescents and high school students replicate the corresponding professional education programs at the bachelor's and master's level, as well as thematic online courses. As the researchers note, higher education programs are not directly applicable for school age, even for high school students, since they are focused on in-depth study of technical solutions, on the consideration of relevant industries. maturing people topics related to data work with data circumstances of life of society.

Consequently, the model of mastering work with data used in vocational education for students of primary and high school should be adapted, brought to less abstract and more simple stages for students. In this regard, in such programs, the emphasis is on lower-level practical skills for working with data in a particular software environment (for example, using Python or R). T. Erickson propose to move through the phases of data exploration in the mode of elementary "steps" (datamoves) - performing specific operations with data, without which no data project can do. Examples include: filtering, grouping data, charting. According to the authors and the experts citing them, this "craft" level of working with data prepares students for a better understanding of more complex material - inductive statistics and the mathematical foundations of machine learning. Another aspect of adapting the “technological” preparation of students to work with data is organizing it on the basis of an integral, albeit local, productive-activity (“project”) task or on the basis of mastering an integral sphere of activity. Here, going through data-moves in mastering data technologies is also an important element, but subordinate to broader educational contexts: import-export of data, checking and fixing data types, encoding categorical variables, renaming-adding-deleting columns, elimination or filling of blank values, search for data outliers, output of descriptive statistics, output of information about a dataset, construction of scatterplots, access to arbitrary elements and filtering of a data set.

Practices that have potential to master working with data as a complex activity.

In this direction, it is possible to single out not so much typical models of educational practices, but the types of basic educational technologies and methods that ensure the recreation of holistic units of activity for students, which are based on working with data.

First of all, it is worth highlighting the conduct of "case-studies" related to the solution of real production situations requiring the use of data analysis. Pupils must analyze real problem conditions, develop their own solutions, and then comprehend them, discuss, compare with the actually made decision, identify the advantages and disadvantages of a real solution and solutions proposed by them. Example: it is necessary to develop a program that allows to regularly calculate how much money needs to be loaded into ATMs in different cities in order to optimize the bank's activities. This practice is largely aimed at developing students' meta-subject and cognitive abilities through meaningful production tasks.

Another group of methods is associated with the organization of practice-oriented learning by students of development tools and programming languages that allow working with data. These include, for example, "Python" and "R". Similar tasks are actually solved

within the framework of other disciplines of the main curriculum, aimed at the general development of information and communication technologies.

Also, methods of working with data are used in the organization of optional work of students. Here it is possible to single out the preparation of speeches by students at specialized scientific and practical conferences as a separate type of educational activity. Moreover, training can be considered both within the framework of areas related to digital technologies, and any other that require analysis and modeling based on large arrays of obviously heterogeneous data, as well as requiring rapid extraction and processing of information from different sources. In this group, we also include additional courses devoted to both theoretical problems and the basic structure of artificial intelligence, and practical tests of the acquired knowledge on a typical ("educational") material that does not have a personally significant status for these specific students, with the development of the necessary sections of the above languages programming "here and now".

A separate place is occupied by "hackathons" and trainings in demanded engineering specialties, during which students first immerse themselves in a problem situation, and then independently decompose it to applied problems that require data analysis tools for their solution.

Practices that ensure the use of data analysis as a means of working with other types of knowledge.

For the organization of subject-cognitive training in the guarantee of an activity-based approach, data analysis is a unique resource, since it allows students to independently select material for study within a specific subject based on open publications and data arrays in the mode of open search and project activities. Students can independently reconstruct and design the subject. Such work should be accompanied by a teacher-mentor, and it should be preceded by the objectification of the supporting, “pivotal” content components of the studied subject as at the same time a system of scientific knowledge - and a specially organized space for the student's culturally appropriate independent action. Such meaningful supports, in our opinion, can be socio-cultural objects associated with the relevant areas of knowledge and practice as phenomena that simultaneously concentrate objective knowledge about reality, hold or realize universally significant meanings, embodied naturally.

An example is educational spaces that allow students to design specific applications based on data and machine learning tools to solve relevant specific research problems. In such cases, students receive a problematic task within the framework of the academic subject; determine what kind of work with large arrays of heterogeneous information will be required to solve it; use machine learning tools, for example, a neural network, which allows solving this problem in an optimal way. It is important that such tasks can be associated not only with physical and mathematical or natural science disciplines, but also with subjects of social science or humanitarian cycles. Under the conditions of mentoring, students are able to develop machine learning models as a tool for independent mastering of subject and practical knowledge, using data analysis methods as a means of cognition.

Self-determination and educational individualization practices

These practices can be divided into 2 main categories: comprehensive educational programs (mainly modularly organized) and individual educational services provided by artificial intelligence.

1. In the case of educational programs, students are offered progressively more complicated tasks of working with data, on the basis of which, the development of specific methods and techniques, the development of basic forms and principles of co-organization, the formation of general ideas about working with data as a tool of economic and socio-management activities. It is important to note that the complication of tasks as the program unfolds is associated not with the inclusion of technologies that are increasingly difficult to understand and apply, but with the consideration of increasingly complex economic, social, managerial, cultural situations - with their objective significance for significant groups of people, with their nature, which complicates the formal description, and, consequently, the collection and structuring of data, with the complication of the context in which the analysis results are used (from a simple description of these results to the development of project proposals based on them). All tasks offered to students are of a trial-practical nature. They are associated not with the reconstruction of culturally fixed, objectified subject knowledge, but with the acquisition of fundamentally new ideas about reality (its phenomena, patterns, etc.), which are initially considered as the basis and support for a specific project action that ensures decision-making and implementation of a productive action. This productive action can be associated with research in the field of fundamental science, with sociological and socio-psychological research, with the study of the political situation. It is important that in any case, the student does not recreate the reality personally for himself, but creates the grounds for his own self-determination in relation to this reality, for taking a socially responsible position. We can confidently say that it is the social positioning of the student that is the key result of such a program. Date literacy becomes its necessary basis and then an appropriated tool for a possible adjustment of the position, obviously based not on assumptions and emotions, but on the processing, verification and correct interpretation of a large amount of information. Such educational results are ensured by the fusion of both humanitarian and technical knowledge within one educational space, where the latter acts as a means of solving the problems of the former.

2. Intelligent tutoring systems (ITS) are considered today as an important support for the mass implementation of an individualized approach to teaching. They use artificial intelligence methods to individually provide educational activities that would best suit their own interests and needs of the student, including based on constant feedback. Some ITS give the student control over their learning to help develop self-regulatory skills; others use pedagogical strategies to optimally advance the student through the curriculum.

An adaptive tutor can include a whole range of artificial intelligence tools, which range from modeling cognitive and emotional states of a student to meta-cognitive support. Examples include Ecolab and AnimalWatch. AnimalWatch's pedagogical design is linked to cultural history theory in psychology. Its developers have formed an operational definition of the "zone of proximal development" and presented a thorough analysis of pedagogical adaptability. This resource implements various types of assistance ("support") and a method for measuring the zone of proximal development. The operational definition of the zone of proximal development shows how to determine it,

how to use supports for the student's independent trial actions within its limits, how to remove these supports in order to maintain the necessary psycho-emotional tone and motivation of the student - to make him experience difficulties, but not overload.

Distinguishing between two categories of practices, we emphasize that when studying the use of educational data to support student self-determination, the category of data collected in the learning process and containing indicators of student activity and behavior, their progress along the educational trajectory, etc. is most often discussed. Such data is used for management purposes and as a measurement tool. However, this approach does not consider the data as didactic material. In relation to the student's self-determination, they act as a means of feedback, rather helping to coordinate and stimulate self-determination, but not to organize it. This task is solved by digital tutors who are arranged as support services. For the organization of self-determination and the impact on its quality, data should act as didactic material, and methods of data analysis - a means of cognition.

To use a “digital tutor”, a student must already go through the first stage of individualization: to form a readiness for the individual character of educational activity, at least to formulate his own interests, and above all, to have substantive reasons to use a “digital tutor” as his own resource. Thus, a digital tutor does not negate the teacher specializing in self-determination, nor does it negate the value of self-determination based on subject matter.

4 Discussion

To work with data can be effectively used in educational practices, it must be based on “authentic” (or “genuine”) data (including selecting appropriate tasks). This definition should be understood as "reliable quantitative and qualitative data extracted from the phenomena of real life." They can be contrasted with "inauthentic" data, which can be artificially generated for demonstration purposes, or can be the result of manipulating data to achieve the intended learning outcome or interpretation.

Authentic data can be collected using a variety of measurement methods and tools (e.g. sensor data), generated by computer models and simulations, or obtained from online repositories and scientific publications. Speaking about authentic data, mean, first of all, data of a natural scientific nature. However, with some assumptions and depending on the content of the curriculum, the teacher may well attribute to them unadapted economic, sociological and other data. At the same time, it is important to note that data literacy develops mainly in those conditions when students not only process and interpret authentic data, but also maintain the context of their use in their own interests, related both to solving a specific educational problem, and in general with life. self-determination. This is where data authenticity becomes especially important. On the one hand, these data are part of an endless array of information about reality and require correlation with the specific goals and interests of the growing up person. On the other hand, they are objective, independent of this young man and therefore presupposing his “counter adaptation” of his goals and interests to the actual state of affairs, represented by these data.

The factor of students' own goal-setting assumes that the data should be not only authentic, but also relevant to the objective interests of students (but not their subjective expectations). It was recorded that authentic data related to subject content, with an understandable context of their creation and in a clear connection with the problems of the real world, is potentially more interesting and attractive for students.

At the same time, the relevance of the data to the initial interests of the students is not enough. It is necessary to provide a special procedure for their "assignment" in the educational process. In our opinion, the optimal solution here is not the direct hermeneutic action of the teacher, who helps the student to correlate his interests with the task, but an objectified, alienated from both the teacher and the student, an educational task that is relevant to the interests of students both in its basic topics and in the requirements for a solution that poses an additional challenge. The action that must be performed within the framework of solving the educational problem must, on the one hand, be relevant and practically significant for the student, and on the other hand, contain "uncertainty", "gap", both in conditions and in interpretation options , as part of possible models and methods of solution. This uncertainty is set both by the basic formulation of the problem and by the use of raw authentic data in the process of solving it. This specificity of the educational program allows not only to ensure high motivation of students for productive activities, but also, first of all, create for students the need to independently set tasks regarding information search, organize this search, reconstruct a hypothetical model of reality and then confirm, refute or correct it in the course of own development, to reconstruct or create from scratch the most effective algorithms and methods of working with the problem.

The logic of the student's activity-based assimilation of data analysis, or the organization of cognitive activity through data analysis, requires a model of co-organization in which the student can alternately, in a trial mode, take different positions within the framework of solving the task. For example, he can perform different roles in accordance with the division of labor within a data project (programmer, data engineer, project manager and subject matter expert). At the same time, each position should involve the student performing the function of a kind of "data producer" - the subject of work with knowledge, registration and presentation of them to beneficiaries. From the point of view of researchers, this approach, which is both multi-positional and supposes the active-transformative nature of each position, ensures that students develop not only skills in working with data, but also a critical attitude to data, as well as students gain experience in overcoming solving obviously non-standard tasks.

5 Conclusions

If we set ourselves the task of not only and not so much the formation of specific near-professional skills in students, but are interested in the development of their self-determination, the formation of functional literacy and the development of metasubject competencies, then we should highlight the requirements for the optimal use of data analysis work that will work to solve similar tasks.

Working with data should take place not just on the material of specific academic subjects or practices, but as a basic tool for mastering these practices, organizing the structure of this mastering, which will allow the student to independently reconstruct the knowledge system and then appropriate it in accordance with their own activity tasks.

Such work with data in the framework of the educational process must be organized in the form of a systemically organized and internally coherent educational program. In turn, the organizing element of this program should be a basic educational task, based on which, students formulate their goals and determine the optimal way of activity. Learning tasks devoted to obtaining specific knowledge or a particular transformation of reality through the use of data analysis become didactic tools here. The basic management principle of such an educational program should obviously be not control over the achievement of formalized indicators, but pedagogical design.

The educational program organizing the development of data analysis by students is formed based on two equal sources:

a) a subject knowledge, which, on the one hand, must be mastered within the framework of the program, and on the other hand, must become a methodology for the student that organizes his work on the use of data analysis as a tool for the reconstruction of the cognitive sphere;
b) a set of technologies, methods, supporting elements of data analysis.

Both of these elements are in a mandatory synthesis: subject knowledge sets the content and structure for working with data, and the data analysis approach itself organizes the development of knowledge in an open, activity-based form.

It is also necessary to highlight the fact that skills in working with information and communication technologies are mandatory, but not the key results of such an educational program. The key integral result is data literacy, understood not as a qualification in working with digital resources, but as the ability to organize one's own life activity based on independently organized work with information, turning it into the basis for making decisions, material for developing these decisions, a source of methods for conducting solutions to life, etc.

All this requires new professional training of teachers for the development of data literacy in modern children and schoolchildren.

ACKNOWLEDGMENTS

The authors thanks the Russian Foundation for Basic Research for the financial support of the grant project №19-29-14016 «Methodology for the analysis of bulk data in education and its integration into training programs for teachers and heads of educational institutions in the logic “Pedagogy based on data”, “Management of education based on data”».

REFERENCES

[1] O.A. Fiofanova. 2020. *Big data analysis in education: methodology and technology*. M .: Business, RANEPA.
[2] M. Schield. 2018. Information literacy, statistical literacy and data literacy. *IASSIST Quarterly* 28, 2, 6-11.
[3] J.R. Carlson and et all. 2011. Determining data information literacy needs: A study of students and research faculty [Electronic resource]. Paper 23. *Libraries Faculty and Staff Scholarship and Research*. Available at: https://docs.lib.purdue.edu/lib_fsdocs/23/ (accessed: 10.04.2020).
[4] E.B. Mandinach and E.S. Gummer. 2013. A systemic view of implementing data literacy in educator preparation. *Educational Researcher* 42, 1, 30-37.
[5] A.F. Wise. 2020. Educating Data Scientists and Data Literate Citizens for a New Generation of Data. *Journal of the Learning Sciences* 29, 1, 165-181.
[6] R. Luckin, W. Holmes, M. Griffiths and L.B. Forcier. 2016. Intelligence Unleashed. An Argument for AI in Education [Electronic resource]. Available at: https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/about-pearson/innovation/open-ideas/Intelligence.
[7] M.K. Kjelvik and E.H. Schultheis, 2019. Getting messy with authentic data: Exploring the potential of using data from scientific research to support student data literacy. *CBE Life Sciences Education* 18, 2, 1-8.
[8] T. Erickson and et all. 2018. Data Moves: one key to data science at school level. *Proceedings of the International Conference on Teaching Statistics (ICOTS-10)*.
[9] A. Cuoco, E.P. Goldenberg and J. Mark. 1997. Habits of Mind: an organizing principle for mathematics curriculum. *Journal of Mathematical Behavior* 15, 4, 375-402.
[10] W. Finzer. 2013. The data science education dilemma [Electronic resource]. Technology Innovations in Statistics Education 7, 2. Available at: https://escholarship.org/uc/item/7gv0q9dc (accessed: 10.04.2020).