STUDIA EKONOMICZNE ● ECONOMIC STUDIES
NR 4 (XCV) 2017

ARTYKUŁY DYSKUSYJNE

Aleksander Żołnierski*

BIG DATA AND REFINING INFORMATION TOOLS IN THE PROCESS OF PUBLIC INTERVENTION PROGRAMMING IN POLAND

(Artykuł nadesłany: 06.02.2018; Zaakceptowany: 02.10.2018)

ABSTRACT

The aim of the article is to describe the use of information refining and big data analysis in the process of programming of public support for R&D activities. The method employed by NCRD in R&D support -related activities is used as an example. Intervention programming is a process dependent on the quality of available information – including its timeliness and usefulness. Usually the information sources used are based on data describing reality with a few years delay and abstracting from the latest trends. New methods of refining information from on-line sources allow to use current data in the programming process (as well as in the evaluation), which adequately describe the rapidly changing reality.

Keywords: public administration, management, information refining, big data.
JEL Classification: H83, M10, O31

INTRODUCTION

Planning the spending of public funds is extremely important from the point of view of national strategy and should reflect the objectives of economic policy based on long-term development trends. Effective and efficient programming of

* Institute of Economics, Polish Academy of Sciences; e-mail: aleksander-zolnierski@aleksander-zolnierski.eu

support for research and innovation activities must be based on two assumptions: reliable and up-to-date data, and transparency of activities and independence from current political objectives. When implementing the strategic objectives of economic policy, one should bear in mind the separation from political pressures. In practice, this may be extremely difficult. It is worth repeating after Schumpeter (1976/2003: 262) that "the typical citizen drops down to a lower level of mental performance as soon as he enters the political field".

From the point of view of the organization in which intervention is planned, the transparency of the process and elimination of organizational pathologies, especially nepotism, require attention. A negative effect of nepotism is the "privatization" of some processes in the organization, where activities, including those resulting not only from the current functions of the system, but also planning and creating strategies are subject to the particular interests of small, informal and usually very hermetic interest groups. In order to reduce these negative phenomena it is necessary to use clear and transparent procedures, but also tools (including IT tools). All this builds unique strategic knowledge. In commercial organizations, knowledge is difficult to copy by the competition, but for each organization it is a unique, strategic resource (Żołnierski, 2015: 103).

Support programming uses knowledge management instruments; both in the soft layer (human capital, social capital, cultural capital) and hard layer (databases, IT, etc.). It is important for the programming process to be based on systematic research of the environment and the accumulation of knowledge. Strategic change takes the form of an increase in the importance of organized research, the implementation of which is often supported by external institutions – entities from the science sector or specialized R&D entities. The role of interdisciplinary teams involving external experts is becoming increasingly important (Janasz, 1999: 37). Over time, these teams must acquire new competences in the use of advanced quantitative methods using IT (Ohlhorst, 2015: 29–30). Data-intensive methods based on IT tools are excellent complements to existing research and analytical methods. This is a way to discover potentially new relationships and improve theory (Chai, Shih, 2017). Monitoring methods which are necessary for programming support for public funds must come out from the organization's mission. Knowledge management processes begin to define the identity and basic competencies of the organization. As with any other organization, those involved in the intervention process have their future depending on the environment, therefore it is necessary to answer the question, what are the dominant development trends (Demecki, Żukowski, 2010). This question defines the scope and method of acquiring knowledge about the organization's environment, and consequently translates into the formulation of its strategic aims.

The issue of timeliness, availability and reliability of data, which were used in the programming as an essential decision-making process so far, is characterized by a relatively long delay to the phenomena or problem it described. For example, the feasibility studies for sectoral programs submitted in 2016 were most often based on statistical data preceding 2014 and for this reason did not fully

reflected the actual condition and potential of the sectors. In addition, there is a lack of objective and up-to-date information on the current and planned activities of organizations implementing research and development.

The aim of the article is to describe new ways of obtaining reliable and objective information for the programming process.

Developing methods of data collection and analysis, including in particular information refining and big data analysis meet the needs of the programming process in terms of obtaining up-to-date and objective information. The use of Big Data analysis allows for more effective monitoring of milieu of the economic sectors. The data obtained in the process of information refining are also independent of the direct influence of ad hoc political interests.

The method I describe is the first application of information refining and Big Data analysis in the programming process in Poland and is one of the first applications of these methods by public administration in the world.

The article summarizes a research project in which the possibilities of information refining and Big Data analysis were used to analyze selected research issues within the sector programs of NCRD. The results were applied in the assessment of the effectiveness of the analyzed programs.

New methods of monitoring the environment, especially those using artificial intelligence (BI), business intelligence systems, Big Data and information refining, are also reflected in the case of programming processes of R&D support (Richards, 2017) and begin to constitute an element of the knowledge management system. The efficiency of management systems depends on the ability to "create, transfer, pool, integrate and exploit knowledge resources". (Frishammar, Richtnér, 2008). The new methods are therefore an assumption for a strategic change in the acquisition and analysis of data from the environment[1]. However, it should be remembered that Big Data is not a panacea for all issues related to the analytics of large data sets. Big Data does not eliminate the need for intuition and creativity (Hayashi, Winter 2014). It is most important, whether the organization is doing enough to develop the skills and competences to achieve the strategic aims. The affirmative answer is also an indicator of the organization's ability to activate a powerful driver of competitive advantage (Ben-Hur, Jaworski, Gray, 2015).

New way of monitoring the milieu consisting of refining information method and Big Data analysis was applied to the identification and exploration of network resources in search of keywords and phrases on the construction materials technologies. The collected data were cleansed and lemmatized and then analyzed with the use of Big Data tools. At this stage, data were structured and subjected of the statistical analysis and then the desired information was obtained.

[1] Author refers to the concept of the project, the description of which is in the further part of the study. The concept of the project was presented to the leadership of the National Center for Research and Development in October 2015, the project was launched in mid-2017. The aim of the project was first of all to develop a solution enabling acquisition and analysis of information from various large data sources (Big Data) covering the Research, Development and Innovation (R&D&I) issues.

The rest of the paper is organized as follows. Section 1 reviews the assumptions of the process of programming the support for R&D activity in Poland, Section 2 presents new, mainly unstructured, data sources, and Section 3 – a description of information refining methods and Big Data in the context of support programming and an example illustrating the method used in the study of construction materials issues (Section 4). The paper concludes with a synthetic summary and a description of the most important conclusions.

1. ASSUMPTIONS TO THE PROGRAMMING OF THE PROCESS OF SUPPORT TO R&D ACTIVITY

Programming of support for R&D activity should include purposeful operations based on the principles of communication rationality. The measure must refer to an objective, social and subjective world, it must also be characterized by normative rightness (Habermas, 1998: 171). Such approach in the programming process of support for R&D activity consists, among others, in integrating the scientific and business environment around the issues of support. It is also important to use adequate, current and reliably processed data. There are many good practices scoping of integration of scientific and business communities around common strategic objectives co-defined by the administration,. For instance, the integration of professional communities at the level of voivodships was facilitated by work on regional strategies. In this case, the process of work on strategies had a significant impact on the creation and development of relations and improvement of competences (Gorzelak, Jałowiecki, 2001: 58–59). A good example of building cooperation and developing social capital necessary for the creation of efficient institutions supporting R&D was also the long-term strategy and the report Poland 2030 (Ministry of Administration and Digitalization, 2011: 60).

Monitoring a dynamically changing environment, especially in the field of R&D, is a particularly difficult task. Statistics covering R&D issues face a number of problems which make analyses not fully reliable, and reliable data may not fully reflect reality. This is due to several reasons, some of which concern the issue of reporting to the Central Statistical Office (GUS) and the reliability of data (both in terms of their scope and "quality"). Difficulties in obtaining reliable data at the time when they are necessary concern many processes (including the management of R&D support activities). These include difficulties in comparing data both internationally and interregionally, and significant delay in time (which often makes it impossible to reliably compare economic data, influencing the programming process of interventions) as well.

Monitoring of R&D activity must go hand in hand with cyclical research of demand for innovations. However, it remains an open question to what extent entrepreneurs should be involved in the process of defining strategic aims of innovation support. This issue is particularly important when we treat commercial companies only as "distributors" of innovations which are "produced" by the

science sector and should be introduced to the market. Such a narrow understanding of the role of business, limiting the possibilities of using in practice solutions such as open innovation or social capital, limit the potential effectiveness of programmed support. On the other hand, the demand of commercial organizations for the results of R&D (in the future) is based on the current demand for products and services created by such organizations. Planning of activities in a business sector is mostly based on extrapolation of the present conditions (in the scope of both demand and the potential to satisfy it). From the point of view of the R&D support programming process, a greater integration of the innovation process is necessary (De Moor, Berte, Marez, Joseph, Deryckere, Martens, 2010: 51).

2. NEW DATA SOURCES

The complexity of the described processes concerns the growing wealth of data for analysis. The analysis process itself is becoming more and more complex. The large amount of ambient data makes it difficult for the analytical process to focus on the need for optimal information for the programming process. On the other hand, at each stage of the innovation process, the organization generates information in various forms in virtual space (including, above all, the Internet). From the point of view of monitoring the environment, an important task is to identify this information, its refining and analysis. In order to effective R&D support, it is therefore necessary to identify and exploit data sources which, in a similar way, show data covering the behavior of innovative entities at different stages of the innovation process. From this point of view, it is also necessary to identify the processes of exchange of knowledge of innovative organizations with the environment; when we know the way in which organizations search for information which is necessary from the point of view of their processes, when we are able to identify this information, then the effect of their analysis will bring knowledge about the implemented and planned processes involving R&D activity (Stephens-Davidowitz, 2017).

The exchange of knowledge, both within the company and with its environment, is influenced by trust, values and norms (OECD, 2006, p.82). An important factor is the use of IT tools by the innovator. Monitoring of sources, where the innovator leaves traces indicating both the technologies and solutions he is looking for, as well as) the knowledge developed within the organization is a key element of the modern programming process.

An example of a practical attempt to describe current and future trends important for innovation processes is the use of Big Data and data refining within a joint project implemented by the University of Warsaw and National Centre for Research and Development. The applied solution may initiate a systemic change in the area of monitoring technological and scientific environment. The tool enables advanced analysis of structured and unstructured data contained in

available databases (relational, hierarchical, network, object-oriented, etc.), file sets, portals, websites and Internet forums, streaming transmissions and other sources on the basis of set-up parameters (in particular: keywords and semantically similar: model images, model fragments of sound recordings, including speech recording, fields and scientific disciplines, including newly created multidisciplinary areas, etc.). The specific objective is to create an ecosystem of interactive scientific, scientific-technical and business information on technologies, based on the sequential analysis of structured and unstructured data available in distributed digital repositories for science, education and an open knowledge society, of which analytical methods and tools are an integral part:

- identification, on the basis of set parameters, of large data sources accessible via the Internet,
- exploration of identified large, variable and diverse sets of data,
- preparation of analytical tools,
- acquisition of data and their analysis aimed at forecasting trends in R&D and innovation activity in Poland in selected technologies and research problems,
- data processing, analysis and visualization of forecasted trends will lead to the acquisition of new knowledge on the status of selected aspects of R&D in Poland.

The project includes data sources on R&D and innovation processes contained in identified databases, unstructured sources, websites, social networking sites, forums, information portals, specialist sources of scientific and technical information, commercial and public data collections, as well as the necessary methodologies for effective analysis of the acquired data. The information obtained on the basis of information collected during the project implementation is characterized by the following properties:

- is independent from the observer (objective),
- shows a synergy feature,
- is diverse,
- is an inexhaustible resource,
- can be reproduced and transferred in time and space,
- can be processed without causing wear and tear.

3. METHODS OF INFORMATION REFINING AND BIG DATA IN PROGRAMMING

The progress of technological changes (especially those based on new technology) forces the search for new solutions in the organization's environment. These solutions concern both access to technologies and management techniques, as well as access to new sources of information and data. Founding the strategy of

an organization supporting R&D activities on external base becomes an important part of a long-term development strategy using knowledge as a key capital. Systematic scanning of available technologies and ideas in the environment has become a strategically important activity of the organization. This is especially due to the growing technological complexity of products and business opportunities related to new technologies (Enkel et al.2009). Thoroughly new opportunities for an organization supporting R&D activities are offered by the use of information refining and Big Data methods.

Data in Big Data come both from traditional databases, e.g. functioning in enterprises and public institutions, which contain so-called internal data, and from others. The data have both defined and unspecified structure, which makes it difficult to distribute and process them using the available IT infrastructure (architecture and analytical tools) and calculation methods. The proper use of data has always been a well-known and very important phenomenon. Big Data, on the other hand, brings much greater possibilities of analysis, faster, more accurate and using data from many sources. Therefore, Big Data is a process of using data, not the mere fact of collecting or downloading it. In the Big Data model, the scope of data collection is extended by data rejected in the BI model due to the lack of structure of data sets and supplemented by additional sources, potentially not directly related to the selected stages of the R&D and innovation process. Algorithms used in the tool are examined and evaluated in terms of their implementation in the designed mechanisms of data processing and inference on the basis of acquired data. The use of BI and artificial intelligence algorithms will enable a much broader analysis of data, processes and factors important in processes, along with continuous self-improvement and improvement of the information provided. The AI algorithms are implemented within Big Data tools, which will make them an integral part of reporting and analysis mechanisms with advanced learning and reasoning mechanisms. Artificial Intelligence provides managers responsible for designing instruments supporting R&D activity in real time with information on R&D activity and detects early warnings of problems before they occur (Moore, 2016). When analysis is based on known algorithms and a lot of data is available, computers give an advantage in terms of information processing and data analysis. Although Artificial Intelligence is developing rapidly, it is important that conceptual and analytical work is carried out with the participation of experts. In many situations, the strongest performance is achieved by people and computers working together (Schoemaker, Tetlock, 2017).

The process that underpins the analysis using Big Data's information refining and analysis tools is a multi-step process. The software architecture consists of three modules: operating system, database system and Big Data robot software. The first step is to identify digital resources available offline and online, including Internet sources, object-oriented databases and streaming media. Identification includes both unstructured and structured data. The identification process is continuous – mainly due to the growing resources of the Internet. The next stage is the collection of identified resources (currently, in August 2018, the collection is

about 40 TB of data). Collection of identified resources is carried out by Big Data robot, which is a specialized ICT system for targeted monitoring and collection of data from indicated websites or other data sources available on the Internet and available offline (Gogołek, Jaruga, 2016, p.109).

The data collected by the system is stored so that it can be reused in the future for various studies. Identified and collected resources are searched for selected keywords – cores and sentiments that define a given keyword. Each information that is registered by the robot contains metadata in the form of a description of the source of information, the date of its publication or download. The Big Data robot system provides the collected data in the form of a CSV file, Excell, and via the database interface for the needs of other systems used in the analysis, e.g. R, Statistica (Gogołek et al., 2017).

Text files collected during the refining of information are further processed. The process of extracting information from text files is complex. This is due to the specificity of languages (e.g. inflectional forms of the Polish language are a potential problem; this problem is solved by means of a lemmatization process). Statistical analysis of the text requires its prior processing (cleaning, normalization). The next stage is the transformation of cores or words in the matrix. Matrices allow statistical calculations. The frequency of occurrence in the analyzed corpus of individual cores or words is calculated. In the case of a survey of multi-period text analysis, time series are created. Time series allow to count the links between cores (keywords) and cores of particular importance (sentiments) for the study. The Pearson correlation is calculated. In the case of series containing outliers, Spearman or Kendall correlations should be used. The following are used for time series analysis: linear or non-linear regression analysis, cluster analysis, correspondence analysis. Then, it is possible to forecast changes in the studied phenomenon (Cetera, Gogołek, 2018, p. 96)

The keywords constitute a statistically verified reference point for selection of sentiments (words, phrases) of evaluative/emotional character, which accompany the studied/evaluated phenomenon. Depending on the subject of the research, sentiments can play the role of thematic contexts, selected by experts and/or automatically. To build a community of experts online is a good solution. This allows creating a system of expert knowledge support without the need to create additional formal structures (Huang, Tafti, Mithas, 2018).

Sentiments can be a combination of two words. Their distance given by the number of characters is defined as a radius. Depending on the subject area of the research project under assessment, the context of sentiments – understood as a statistically significant "distance" – needs to be defined. The result of statistical calculations of the relation between verified sentiments and the phenomenon in time intervals (t_1, t_2, ..., t_n) determines the possibility of prediction on the basis of sentiments – predicators – of the phenomenon at t_{n+1} (improvement or deterioration of the value of quotations/assessments). These calculations – prediction – may use functions taking into account more than one parameter (sentiment), e.g. multiple regression analysis with the use of forecasting the values of

explanatory variables. The collected data allow for the creation of information of high usefulness mainly due to considerations:

- substantive competence, which means that the content of the information is adapted to the issue being the subject of the decision,
- selectivity – only those that are relevant for decision making are provided from potential substantive information,
- the appropriate degree of detail,
- truthfulness, i.e. reflecting facts,
- speed and timeliness,
- continuity and regularity,
- unambiguity,
- assimilability by the recipient.

The main task of information is to create the basis for management, whose functions are planning, organizing and controlling, therefore the task of the information created within the project is to create the basis for both current and future activities.

4. EXAMPLE – CONSTRUCTION MATERIALS TECHNOLOGIES

A comparative analysis of the main structural materials of the 21st century shows that steel and cast iron alloys are at the forefront, both in terms of performance and recyclability at a high level. The Polish steel sector has been included in the European and global market. Although steel is a traditional material, it is necessary to continue research on, among others, multiphase steel technology, steel for special applications hardened cast iron. The development of light metal alloys, mainly aluminum and magnesium, used increasingly as construction material for modern means of transport, is of fundamental importance for the development of the national economy. In the area of construction materials based on polymers, four basic groups can be distinguished: thermoplastics (thermoplastics), chemo- and thermosetting plastics (duroplastics), elastomers and composites with polymeric matrix. All these groups of polymeric materials are produced in Poland, but in insufficient selection and quantity.

Research in the field of structural materials technologies covered by the Techmatstrateg program, managed by the National Center for Research and Development, focused on multiphase steel production technology, special steels and ductile production technology of isothermally hardened cast iron. The technologies of light metal alloys – aluminum and magnesium – were distinguished. In the field of construction materials based on polymers four basic groups were distinguished: thermoplastics (thermoplastics), chemo- and thermosetting plastics (duroplastics), elastomers and composites with polymeric matrix. Moreover, the problems of polymer nanocomposites technology and diffuse nanophase

were distinguished. As part of the study, poles covering the following areas of technology were analyzed:

1. technologies of production of high-strength materials from light and superlight aluminum alloys, magnesium alloys and titanium alloys of structural elements with antibacterial effect from copper and copper alloys (Cu+ systems), refractory oxygen nitride materials for contact with liquid metals and thermal insulating ceramic materials with increased mechanical properties,
2. technologies for manufacturing long structural layered composite products based on powder metallurgy,
3. technologies for manufacturing nanoparticles, nanofibres and polymer nanocomposites and technologies for processing polymeric materials, including giving them the desired properties, e.g. highly hydrophilic or hydrophobic and new generation concretes, concrete and bituminous surfaces, self-cleaning and self-cleaning materials and new materials for building traffic safety devices.

On the basis of the collected empirical material, the following topics for further research were defined:

1. lightweight aluminum alloys,
2. magnesium alloys,
3. titanium alloys,
4. copper alloys,
5. oxygenitriding refractory materials,
6. ceramic insulating materials,
7. powder metallurgy,
8. manufacturing technologies for nanoparticles,
9. manufacturing technologies for nanofibres,
10. manufacturing technologies for polymer nanocomposites,
11. polymer material processing technologies,
12. self-healing material; self-cleaning material,
13. concrete pavements, bituminous,
14. austempered ductile irony; ductile irony,
15. isothermally hardened.

The topics formed the basis for the definition of 20 keywords, adopted as separate research tasks. Table 1 presents the results of calculations of the frequency of occurrence of contractual sentiments defined as attributes (the most common words occurring in the vicinity of the keyword). The environment is measured by the number of 60 letters before and after the keyword.

All available information channels (sources) covering the period from 2017-01-01 to 2018-07-31 were selected for the study. The main objective of each

Table 1. Results of calculations of the multiplication of keywords describing the subject of research

Form of keywords	Frequency
Nanoparticle	161 723
Pavement + sidewalk	23 880
Self-healing	6 758
Polymer and material	5 040
Construction and material	3 635
Magnesium and alloy	3 379
Titanium and alloy	2 706
Polymer and nanocomposite	2 238
Self-cleaning	1 617
Bituminous	1 465
Powder and metallurgy	1 264
Nanofibre	738
Copper and alloy	630
Ductil and iron	438
Concrete and pavement	310
Austemper and iron	128
Lightweight and aluminum	87
Ceramic and insulating	5
Isothermally hardened	0
Oxygenitrid and refractor	0

Source: own elaboration based on project results

study was to assess current changes in the state of innovation in the subject under study and to predict the strength and direction of changes in this state in the nearest future. The period of 18 months was adopted as a statistically reliable basis for the assessment of current changes in innovation and the prediction of these changes. It is found that the study of a longer period does not provide statistically significant information that can be the basis for obtaining more accurate results, measured by the accuracy of prediction. As a justification for this statement, the historical data obtained in period t–2 (year) were used at time t of the study, calculating the prediction of change in period t–1 (1–2 months). The result of the comparison of the prediction in period t–1 with the actual changes in time t–1 is statistically insignificant. Predictions are comparable to actual predicted changes. This confirms the earlier claim that omitting data older than 12–18 months is justified. The technologies represented by keywords were, from the point of view of solutions created in Poland. the most important areas of research and implementation; the described picture covered the period

Table 2: Summary of quantitative assessments of the intensity of keyword occurrence and the expected change in the dynamics and direction of popularity changes

Keyword	Frequency	Forecast
Nanoparticle	161 723	↑
Titanium and alloy	2 706	↑
Polymer and nanocomposite	2 238	→↑
Magnesium and alloy	3 379	→↑
Polymer and material	5 040	→↑
Construction and material	3 635	→↑
Powder and metallurgy	1 265	→↑
Self-cleaning	1 617	→↑
Copper and alloy	630	→↑
Lightweight and aluminum	87	→↑
Ceramic and insulating	5	→↑
Nanofibre	738	→↑
Bituminous	1 465	→→
Pavement + sidewalk	23 880	→→
Ductil and iron	438	→→
Self-healing	6 758	→↓
Concrete and pavement	310	↓
Austempered iron	128	↓

Source: own elaboration based on project results.

from January 2017 to July 2018. It is important to examine and analyze the topicality of the defined technologies and research problems. Due to the respective abundance of occurrence, fifteen keywords were analyzed[2]. The analysis showed the intensity and trends in the scope of interest in the keywords and their attributes.

Particular intensity concerns: nanoparticle, titanium and alloy (upward trend), concrete pavement and austempered iron (downward trend).

Above (Tab. 2) there is a list of keywords with their attributes broken down into those attributes for which the forecast indicates an upward trend, for which the forecast indicates a downward trend and those which are in a stagnating state or where an upward or downward trend cannot be determined.

[2] Nanoparticle, pavement, sidewalk, self-healing, polymer and material, construction and material, magnesium and alloy, titanium and alloy, polymer and nanocomposite, self-cleaning, bituminous powder and metallurgy, nanofiber, copper and alloy, ductil and iron, concrete pavement, austempered iron, lightweight and aluminium.

SUMMARY AND CONCLUSIONS

IT, Big Data and information refining tools built during the project were used to analyze previously unused data sources. The presented example of the use of information refining indicates a possible identification of such data sources which, including elements of R&D processes located in dispersed databases, social networking sites, forums, information portals, specialist sources of scientific and technical information, commercial and public data collections, allow for more effective public intervention. The project was an experiment, and its results show a pioneering, innovative solution for the use of Big Data for R&D support management.

The project was based on the following assumption:

- the available sets of data on R&D include information from the past (usually the 'delay' of information is two years; for example, data from EUROSTAT, GUS or other institutions involved in monitoring R&D on a macro- and mesoeconomic scale),
- data covering such a distant past are not optimal from the point of view of assessing the effectiveness of intervention or planning the implementation of new R&D support policy instruments,
- it is already known that a certain range of organizational behaviors related to R&D activity is correlated with a number of signalling factors appearing in the environment of the organization,
- monitoring of these signals, with the use of complex analytical models, allows for obtaining the most up-to-date information on current and planned activities of organizations implementing research and development activities, and the use of Big Data analysis will allow for more effective monitoring of the milieu, macro-environment and individual sectors in the context of projects implemented within the framework of Smart Growth Operational Program 2014–2020.

The research, its methodology and tools, based primarily on Internet data, has a large development potential, which results not only from the growing physical resources of the network, but above all from the increasing activity of its users, who generate more and more data resources. Collection and analysis of this data is possible on the basis of developed technologies, which already allow to collect and process huge amounts of data.

The use of IT tools can be described as:

- access to current data on R&D activity at the scale of economy, regions and industries
- access to data that will make it possible to predict,
- the R&D and innovation activity in the future,
- the demand for human resources involved in R&D and innovation processes,

- the competitiveness of the economy and industries,
- increasing the possibility of carrying out reliable evaluations using counterfactual methods,
- increasing the potential for evidence-based policy making by using data,
- faster, more precise reaction to the situation in the scope of R&D in Poland (both for the needs of scientific and commercial sector, as well as the development of scientific and research potential).

Big Data is already used by many institutions (e.g. in the USA) to predict, among other things, the economic condition of stock market companies, or even the probability of crime (such a program is carried out by the police in Modesto, California) (Zhang, Luna-Reyes, Pardo, Sayogo, 2016). Described project is a pioneer in the use of Big Data methods by the administration in Poland. The use of the applied methodology may be important also outside the programming itself. Tools can be found wherever they are analyzed:

- legal instruments – including acts and regulations introducing specific economic and financial system regulations, technical and organizational requirements, reporting and information obligations,
- financial instruments - including those concerning taxes, fees, allowances, subsidies, preferences, preferential loans, loan guarantees and contributions which would be necessary to achieve the assumed objectives within economic policy,
- institutional instruments – among them: changes in competences and principles of coordination of tasks of state institutions, creation of institutions or enterprises, changes in the monitoring system, creation of information and training systems, which are a necessary element of strategy implementation,
- programming instruments, including proposed development programs or other programs as instruments for the implementation of the national strategy.

The project proves that methods used allow for more effective monitoring of R&D in selected economic sectors. Data obtained in the process of information refining and Big Data analysis, as independent of the direct influence of ad hoc political interests, are a great way to obtain up-to-date and objective information in the programming process and evaluation as well.

REFERENCES

Ben-Hur Sh., Jaworski B., Gray D. (2015), *Aligning Corporate Learning With Strategy*, MIT Sloan Management Review", Fall, 57(1).

Cetera W., Gogołek W. (2018), *Identyfikacja innowacji za pomocą analiz Big Data*, in: K. Opolski, J. Górski, ed., *Innowacyjność polskiej gospodarki: wybrane aspekty*, WNE UW.

Chai S., Shih W. (2017), *Why Big Data Isn't Enough*, "MIT Sloan Management Review", Winter, 58(2).
Demecki W., Żukowski P. (2010), *Budowa strategii jako narzędzia innowacyjnego zarządzania organizacją*, Prace Komisji Geografii Przemysłu, Warszawa-Kraków, nr 15.
Enkel E., Gassmann O., Vanhaverbeke W., Vrande V. van de (2009), *Open innovation in SMEs: trends, motives and management challenges.* "Technovation, Zoetermeer", 29: 423–437.
Frishammar J., Richtnér A., (2008), Editorial, "Int. J. of Technology Intelligence and Planning", 4(3).
Gogołek W., Jaruga D. (2016), *Z badań nad systemem rafinacji sieciowej. Identyfikacja sentymentów*, „Studia Medioznawcze", 4(67): 109.
Gogołek W., Celiński P., Cetera W., Grzegorek J., Jaruga D., Kononiuk T., Kowalik K., Kuźmina D., Opolska-Bielańska A. (2017), „Eksploracja źródeł danych w zakresie działalności B+R+I – dokumentacja projektu", materiał niepublikowany, NCBR, Warszawa.
Gorzelak G., Jałowiecki B. (2001), *Strategie rozwoju regionalnego województw: Próba oceny*, „Studia Regionalne i Lokalne", 1(5): 58–59.
Habermas J. (1998), *On the Pragmatics of Communication*, MIT Press, Cambridge, Mass.
Hayashi A.M. (2014), *Thriving in a Big Data World*, "MIT Sloan Management Review", Winter, 55(2).
Huang P., Tafti A., Mithas S. (2018), *The Secret to Successful Knowledge Seeding*, "MIT Sloan Management Review", Spring, 59(3).
Janasz W. (1999), *Modele strategicznego zarządzania innowacjami w przedsiębiorstwie*, Informa, Szczecin.
Ministerstwo Administracji i Cyfryzacji, *Polska 2030 Trzecia fala nowoczesności, Kreatywność indywidualna i innowacyjna gospodarka* (wersja robocza), kluczowe decyzje.
Moor K. De, Berte K., Marez L. De, Joseph W., Deryckere T., Martens L. (2010), *User-driven innovation? Challenges of user involvement in future technology analysis*, "Science and Public Policy", February, 37(1).
Moore A.W. (2016), *Predicting a Future Where the Future is Routinely Predicted*, "MIT Sloan Management Review", Fall, 58(1).
OECD/Wspólnoty Europejskie (2006), *Podręcznik Oslo – zasady gromadzenia i interpretacji danych dotyczących innowacj*i, wyd. polskie: Ministerstwo Nauki i Szkolnictwa Wyższego, Warszawa.
Ohlhorst F.J. (2015), *Big Data Analytics: Turning Big Data into Big Money*, John Wiley&Sons, Inc.
Richards G. (2017), *Big Data and Analytics Applications in Government: Current Practices and Future Opportunities*, CRC Press.
Schoemaker P.J.H., Tetlock Ph.E. (2017), *Building a More Intelligent Enterprise*, "MIT Sloan Management Review", Spring, 58(3).
Schumpeter J.A. (2003), *Capitalism, Socialism and Democracy*, George Allen&Unwin (Publishers) Ltd, 1976, Edition published in the Taylor & Francise-Library.
Stephens-Davidowitz S. (2017), *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*, Dey.

Zhang J., Luna-Reyes L.F., Pardo Th.A., Sayogo D.S. (2016), *Information, Models, and Sustainability, Policy Informatics in the Age of Big Data and Open Government*, Springer.

Żołnierski A. (2015), *Nieformalne źródła informacji w działalności gospodarczej*, w: W. Cetera, K. Kowalik (ed.), *Logistyka i administrowanie w mediach*, Instytut Dziennikarstwa Uniwersytetu Warszawskiego.

BIG DATA I RAFINACJA INFORMACJI W PROCESIE PROGRAMOWANIA INTERWENCJI PUBLICZNEJ W POLSCE

STRESZCZENIE

Celem artykułu jest prezentacja koncepcji wykorzystania rafinacji informacji i metod big data w pozyskiwaniu nowych źródeł informacji w procesie programowania wsparcia ze środków publicznych. Przykładem jest zastosowanie tych metod w działalności NCBR. W artykule zamieszczony jest także przykład praktycznego zastosowania opisanych metod w kontekście działalności B+R. Programowanie interwencji jest procesem zależnym od jakości dostępnej informacji – w tym od jej aktualności i użyteczności. Zazwyczaj wykorzystywane źródła informacji opierają się na danych opisujących rzeczywistość z kilkuletnim opóźnieniem i abstrahujących od najnowszych trendów. Dostępne metody rafinacji informacji ze źródeł on-line pozwalają, aby w procesie programowania (ale także w ewaluacji) wykorzystywać dane aktualne i adekwatnie opisujące szybko zmieniającą się rzeczywistość.

Słowa kluczowe: administracja publiczna, zarządzanie, rafinacja informacji, big data.

Klasyfikacja JEL: H83, M10, O31