


CASE STUDY

Open Access



# Potential for the use of large unstructured data resources by public innovation support institutions

Wiesław Cetera<sup>1\*</sup> , Włodzimierz Gogolek<sup>2</sup>, Aleksander Żołnierski<sup>3</sup> and Dariusz Jaruga<sup>1</sup>

\*Correspondence:

w.cetera@uw.edu.pl

<sup>1</sup> The Faculty of Political Science and International Studies, Krakowskie Przedmieście 26/28, 00-928 Warsaw, Poland  
Full list of author information is available at the end of the article

## Abstract

Effective programming of research and development (R&D) support, adjusted to the actual potential of beneficiaries, requires the use of modern analytical tools. An efficient R&D support system requires up-to-date data on technological trends, ongoing (and planning) research, market needs and developing innovation. The most popular programming methods were based on the analysis of data with a 4 to 5-year time delay until recently. Having described the method of refining information from unstructured data, we explore how to make it possible not only to solve the issue of up-to-date data but to identify of the latest trends in R&D activities.

The analytical tools we describe were already fully functional in 2018 and are constantly being improved. The article presents the potential of one tool that can be applied in public support institutions. Methods of identifying and diagnosing technology trends are presented within the case study of the *electric car* technology trend. The presented case study shows the effectiveness of the method we developed for identifying and diagnosing areas requiring support from public funds. Public institutions, including public institutions supporting R&D and innovation processes, can apply tools that allow an increase in the quality of public support programmes offered, but also beneficial for the quality of strategic resources management within the institution itself. The comparison of the predictions made by the described tools with the classifications made by experts, the former are more accurate and precise. Moreover, the results of the analyses performed by the presented model are not influenced by distorting factors—fads, trends, political pressures, or processes with an unidentified, non-substantive background. It should be emphasized that the accuracy of the whole model is 0.84. The described tools and methods are already directly applicable in many areas related to the support of R&D activity worldwide. The article presents a solution that effectively enables the management of more precise programmes supporting innovative activities used for the first time in Poland. It is also one of the first uses of these methods by public administration in the world. Our approach not only strengthens improved adjustment of the support offered for R&D activity, but also makes it possible to apply and improve management methods in public institutions.

**Keywords:** Big Data, Information refining, Information technologies management, Research and development management, Research and development support programming, Data management, Business statistics, Innovation

**JEL Classification:** C810, H110, O320

## 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 support for research and innovation activities must be based on two foundations: 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 need for isolation from political pressures. In practice, this may be extremely difficult. It is worth repeating after Schumpeter [1] 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 information technologies (IT) tools]. This all builds unique strategic knowledge. In commercial organizations, knowledge is difficult for the competition to copy, but for each organization it is a unique, strategic resource [2].

Support programming uses knowledge management instruments, both in the soft layer (human capital, social capital, cultural capital) and in the 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 research and development (R&D) entities. The role of interdisciplinary teams involving external experts is becoming increasingly important. Over time, these teams must acquire new competences in the use of advanced quantitative methods using IT [3]. 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 [4].

The transformation of unstructured data into information that is a useful repository of knowledge is facilitated by, for example, automatic categorization and information extraction [5].

Monitoring methods, which are necessary for programming support for public funds, must emerge from the organization's mission. Knowledge management processes begin to define the identity and basic competencies of the organization. As with any other organization, the future of those involved in the intervention process is dependent on the environment, therefore it is necessary to answer the question as to what the dominant development trends are. 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 the 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 reflect 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.

Typically, the use of data in the programming process in year  $X$  is based on the available statistical data that describe the situation in year  $X - 2$ , and these data are used by policy makers to program supporting instruments for years  $X + 2$ . To solve future R&D dilemmas is based on the analysis of data with a 4 to 5-year time delay. The issue related to the time delay can be solved by means of big data analyses of unstructured but up-to-date data.

A similar problem of time delays is much wider and concerns many areas of life, and building forecasting models and improving data accuracy are a challenge that we deal with in many cases when demand forecasting is necessary [6].

Developing methods of data collection and analysis, including in particular information refining and Big Data analysis, meets the needs of the programming process in terms of obtaining up-to-date and objective information. The use of Big Data analysis allows more effective monitoring of the 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.

Every time we seek to develop a knowledge base from unstructured data sets, we face the challenge of discovery. In this process, it is useful to use many methods that are already known and used by many—e.g. rough set theory, theory of approximation and fuzzy rough sets theory [7].

A novelty in our approach is the use of the described method and information refining and Big Data analysis in the programming process for the first time in Poland. It is also one of the first uses of these methods by public administration in the world.

New methods of monitoring the environment, especially those using artificial intelligence (AI), business intelligence systems, and Big Data and information refining, are also reflected in the case of programming processes of R&D support [8] 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” [9]. The new methods are therefore an foundation for a strategic change in the acquisition and analysis of data from the environment. 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 [10]. What is most important is whether the organization is doing enough to develop the skills and competences to achieve the strategic aims. An affirmative answer is also an indicator of the organization's ability to activate a powerful driver of competitive advantage [11].

More and more examples of techniques and methods related to the analysis of unstructured data can be found in the literature. Unsupervised text processing methods are used [12], and the challenges of creating databases and analyzing data from unstructured text from scientific publications are a key issue when it comes to biomedical data. For effective analysis of this type of data and their integration with structured data, many complex IT systems are being created [13]. In the case of biomedical data, it is important not only to improve the quality of data analysis, but also to protect the data and limit access to it [14]. The analytical tools we describe require adequate computing power, but this allows the time of data processing to be shortened. The tools we use are comparable to others that have recently been developed and implemented [15], but our solution was already fully functional in 2018 and is constantly being improved.

One should also keep in mind the fact that unstructured data analysis and the database tools used require effective data management methods. Similar methods are used in the case of databases based on IoT sources [16].

A new way of monitoring the milieu consisting of refining information methods and Big Data analysis was applied to the identification and exploration of network resources in search of keywords and phrases on 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 to statistical analysis and then the desired information was obtained.

The method we describe is one of many that are successfully used in the world for more effective forecasting of technological trends. The development of methods, and hence the development of text analytics, will continue to grow and develop. The value of the analytical market in this regard is predicted to grow to USD 16.6 billion by 2025, from USD 5.9 billion in 2020, and the compound annual growth rate (CAGR) is projected at 23% for the period 2020–2025 [17].

### Case description

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 [18]. Such an approach in the programming process of support for R&D activity consists, among other things, in integrating the scientific and business environment around the issues of support. It is also important to use adequate, current and reliably processed data. Integration of scientific and business communities and focusing on common strategic objectives co-defined by the administration has many good examples. 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. A good example of building cooperation and developing the social capital necessary for the creation of efficient institutions supporting R&D was also the long-term strategy of development in Poland.

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 time delay (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 into demand for innovations. However, it remains an open question to what extent entrepreneurs should be involved in the process of defining the 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 solutions such as open innovation or social capital in practice, limits 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 from 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 [19].

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, refining it and analysis of it. For 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 the 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 yield knowledge about the implemented and planned processes involving R&D activity [20].

The exchange of knowledge, both within the company and with its environment, is influenced by trust, values and norms. An important factor is the use of IT tools by the innovator. Monitoring of sources and the knowledge developed within the organization are a key factors of the modern programming process as well as to indicate the innovators’ traces leaving indicating both the technologies and solutions they are looking for.

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 the

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 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 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.

The aim of this article is to describe new ways of obtaining reliable and objective information for the strategic management of support for the national innovation system. It presents partial results of a project on monitoring technology trends. The result of monitoring is, but not only, the identification of technology trends and quantitative assessment of their dynamics of change and prediction.

The system created in the project makes it possible to distinguish strategic/dedicated sources of information for identifying technology trends. The following tools were developed:

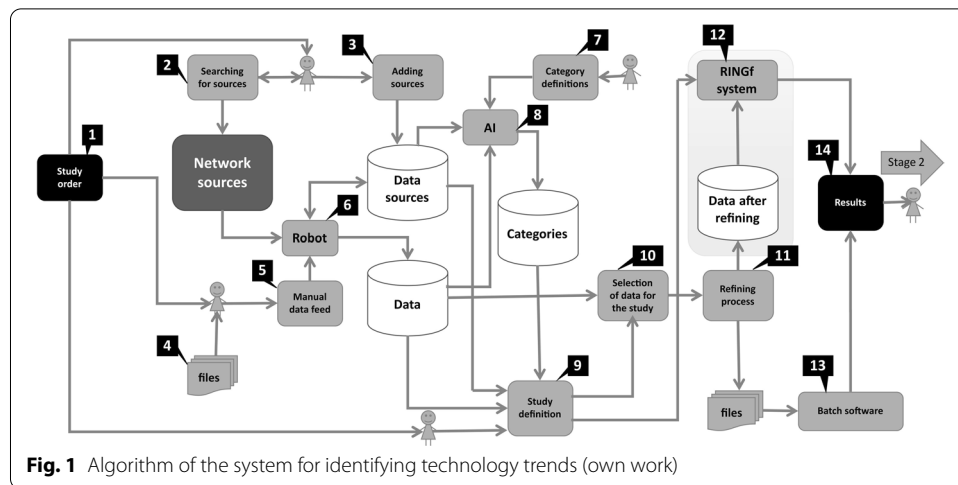
- for automatic collection of source materials,
- for statistical processing,

- for quantitative analysis of historical, current and predictive data,
- systems for visualisation and description of results.

Procedures were defined for identifying dedicated sources and highlighting technology trends in selected areas of fields. The system includes the Analytical Warehouse containing dedicated source materials and software for feeding the warehouse new data. The physical structure of the tool has a layered architecture, and the modular design ensures openness and scalability. The system is characterized by technological neutrality of the mutual information mechanisms and allows modernization of selected elements in the process of expansion. The data update of the Analytical Warehouse is provided by the Big Data Robot, a specialized IT system for targeted monitoring and collection of data from designated websites and offline sources, also in the form of digital text images, e.g. image documents, books or archived newspapers. The system consists of a number of simultaneously operating robots (agents), each of which, in specific and defined time units, interacts with a monitored information source—e.g. an information service. Selective retrieval software allows unnecessary information to be filtered out and a file prepared in a predetermined format to perform computations. It has a built-in function to convert text data saved in HTML format to plain text in UTF-8 format, and a function to convert multi-line text to a single line with the removal of redundant spaces and tabs. The selection of retrieved data reduces the randomness of collecting source material from vast resources. The system ensures that source materials are integrated and standardized, cleaned and classified, and collected in appropriate thematic databases. The source material databases, cleared of unnecessary characters, are the direct “raw material” for performing technology trend identification operations. Assigning materials to field-specific sections (material dedication) and cleaning improve the efficiency and speed of technology trend monitoring functions. These functionalities allow controlled integration of new material, obtained from different sources, with previously collected resources, corresponding to the field under study.

In identifying data sources, resource search strategies were defined and categorised. Strategies are selectively applied depending on the research problem under analysis—the field, the availability of sources and the information and search language used in the source. Specification of tools for information retrieval has been developed, including traditional and specialized search engines, scientific multi-search engines, search engines for documents saved in a specific format. The available digital resources were systematized and the access points to information from different sources were specified: social media aggregators, mashup tools, quality-controlled catalogues, databases aggregating open scientific resources, and discovery systems.

The system is fed with resources related to innovations and technologies in the given field. A mechanism for automatic document classification was developed based on machine learning methods—supervised learning algorithms and natural language processing algorithms. This mechanism will be used each time a new database is fed with new data to divide the database resources according to the defined source categories. A mechanism was also created to classify documents according to any classification, as an additional way of dividing database resources.



The system allows specification of technology trends concerning any area of the world and enables analysis of the dynamics, directions and prediction of changes of identified threads as well as cluster analysis.

The research uses tools to identify trends—elements of artificial intelligence, in particular semantic analysis of Big Data using advanced methods of natural language processing. The result of the analysis is a set of structured data characterizing each layer of the analysed text: the token, syntactic group, sentence, section and the whole text.

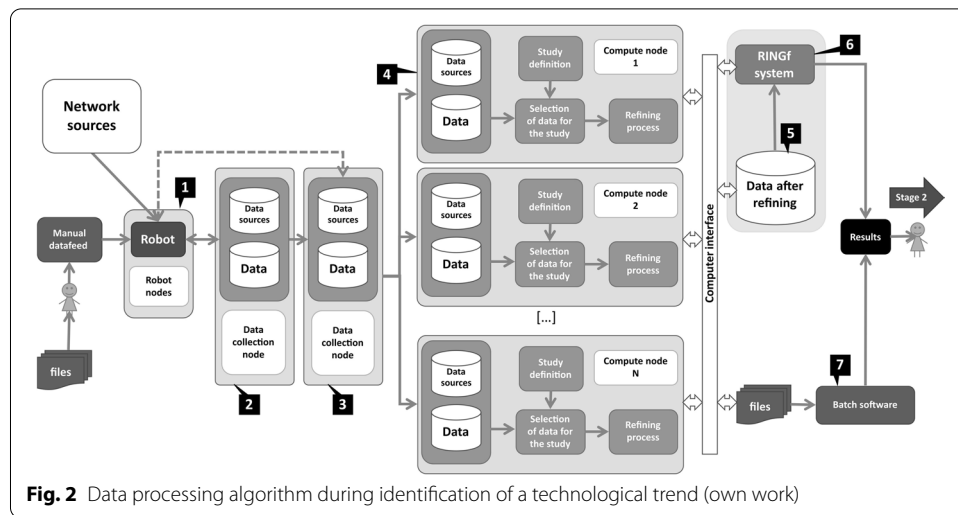
The data refining method consists of cleaning and refining the information stored in large information resources. Information created on the basis of data obtained in the refining process will allow identification and analysis: quantitative description of the dynamics of changes in technology trends in the dimensions of scientific information sources, administration and business sources. The relevance of this analysis lies in the original method of analysing the content related to the post (citation). Reflecting the content of the source materials containing the posts are the post attributes—words occurring in the vicinity of the post. The quantitative and qualitative description of the dynamics of changes in the frequency of occurrence of attributes is an equivalent, with a similar description of the pole, and part of the image of the pole. In the adopted solution, attributes are exogenous variables.

The algorithm of the system for identifying technology trends is shown in Figs. 1 and 2.

Key:

1. Start—information necessary to determine the scope of the study. Includes, but is not limited to, the hypothesis, purpose, and assumptions. Due to the nature of the study, important input parameters determining, among other things, the assumptions are as follows: the topic of the study, the time interval of the research, key words defining the subject of the study, recommended data sources and their categories. Selection of these elements is closely connected with the subject of the study and for each study an individual analysis is required in the scope of necessary input data, selection of sources, data collection, performed analyses, etc.





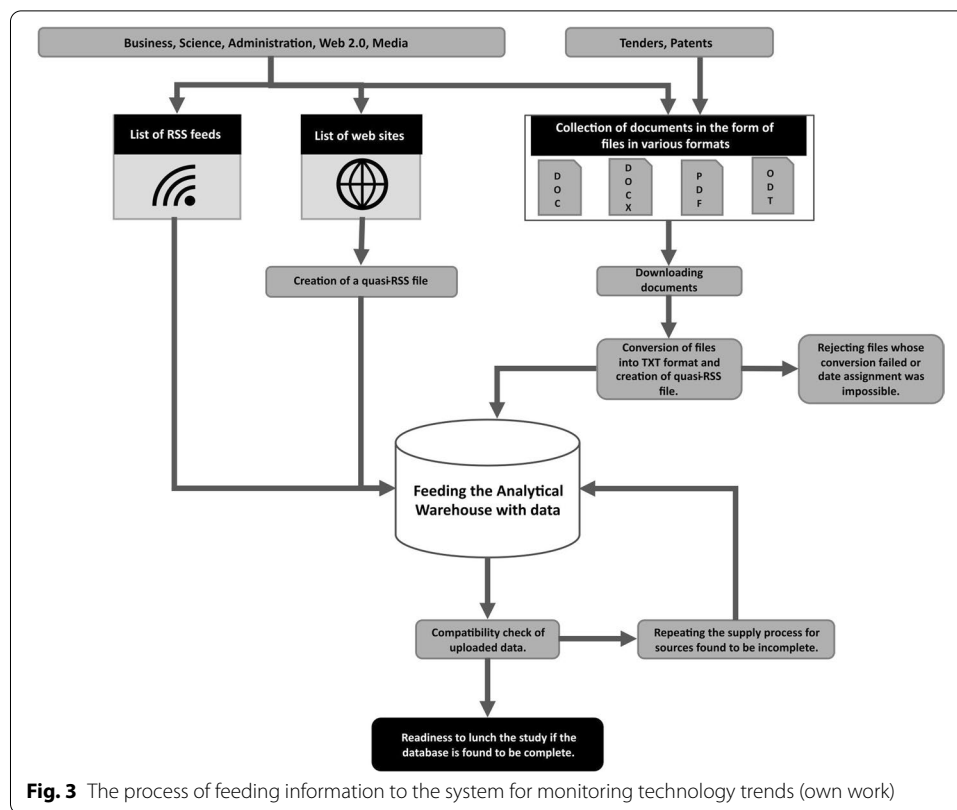
2. Searching for sources—a process conducted by the researcher whose aim is to find appropriate sources of data according to the subject of the study. Searching for sources is done manually each time using the procedures described in this paper. At this stage, a division is made between sources that can be collected fully automatically by the robot and those that require semi-automatic or manual processing.
3. Adding sources—actions performed from the administration panel, by means of which the researcher introduces sources of information to the source database. Based on the data entered, the robot proceeds to collect data. This process is fully automatic, and no direct manual operation is required.
4. Files—represent any type of material that is not available in web sources. This includes the client's own material, which, for various reasons, cannot be accessed from the web. In a sense, we can refer to these materials as offline materials. These files include materials digitised by the researcher or the client as well as any other materials which require non-standard processing, e.g. the conversion of a multimedia file (image, audio) into text.
5. Manual data feeding—this is a functionality of the system that allows feeding data to the robot base, which the robot cannot handle automatically due to the unavailability of this data in the network. The manual data feed of the robot database is complementary to the automatic feed mechanism.
6. Robot—a fully automated system for online data collection.
7. Category definition—based on the information provided in the order, the researcher defines categories of data for a given study. The defined categories are processed by AI algorithms in such a way that individual data sources are assigned to specific categories. The AI mechanism for category assignment is described in the Report.
8. AI—a separate fully automated AI engine built on top of TensorFlow.
9. Study definition—the activities that define a study in the system are derived from the order and processes described above. When defining a study, the researcher determines which data sources, which categories, and on which data the study will be conducted. The parameters defined in this section are the input configuration for the test data selection (10) and the refining process (11).

10. Selection of data for the study—an automated process designed to identify specific documents to be retrieved for the study—the refining process. The input data for the selection are the boundary parameters given during the definition of the study based on the order information.
11. Refining process—a fully automatic process implemented depending on the needs of one or more computing nodes. As a result of the refining process, the system generates files with a database structure or saves data in the ‘Data after refining’ database.
12. Information Refining System (RINGf system)—a fully autonomous system processing data after the refining process into human readable form. The data generated by the system in the form of raw results are the basis for analysis, on the basis of which reports are prepared in the second stage.
13. Batch software—includes proprietary software that performs data analysis after refining in an automated manner. Batch software, in addition to previously defined scenarios, can implement custom data processing algorithms.
14. Results—are a set of files that is the basic raw material needed to prepare the report.

**Key:**

1. The robot is a fully automated data collection system operating 24 h a day.
2. It performs collection of data, which it stores in a database. The robot is run on several nodes in such a way as to ensure uninterrupted collection operation. Depending on the situation, the robot saved data in the primary database (2), and in case of failure or unavailability of the primary server, data is saved in the backup database (3).
3. Primary database server—This is the main server that accepts data. Its purpose is to save data to the database and perform replication to the backup server (3). No computation or information retrieval tasks are performed on the basic database server.
4. Backup database server—its main task is to accept a replica of data from the primary server and make the selected data available to servers acting as compute nodes.
5. Compute nodes—these are powerful servers whose primary task is to perform computations. Each compute node has a replica of the database taken from a standby server. On servers, which are compute nodes, the tasks of data selection, data refinement processes, and computations are performed. More compute nodes can be added to the system as needed. On the compute nodes, computations are performed in parallel. The information broker is responsible for the division of tasks.
6. Data that are the product of refining are loaded into a dedicated database, which is used by the RINGf system or exported to files with a database structure.
7. RINGf system—is the previously mentioned fully autonomous system processing data after the refining process into human readable form.
8. Batch software—is the proprietary software described earlier for post-refining data analysis.

The modular design of the system ensures its open architecture and scalability, with respect to functionalities that are parametrizable depending on the subject of the study, the number of available information sources, the number of source materials. Scalability and openness also apply to the infrastructure of the system itself. This



enables the use of distributed task processing. Processing is done in three compute nodes (strong servers, mainly thanks to super-fast SSD/M2 drives) and ten physical task machines.

The choice of a particular tool, resource or source is always dictated by the specifics of the research area under analysis (Fig. 3). This, in turn, accounts for the fact that each search is heuristic in nature and precludes the use of an invariant pattern that might apply to each search/identification of information sources conducted. The identification of information sources, through various tools (search engines, discovery services, databases, etc.), uses methods that relate to the theory and practice of using appropriate information search strategies. Search instructions are formulated that correspond to the assumptions of several strategies (appropriate to a variety of situations, handling methods, but also sources characterized by a different level of use of the controlled vocabulary). The application of specific search strategies is always related to the research problem analysed, the availability of sources (e.g. specialist sources), the available search tools, the information and search language used in a given source, etc. Consequently, the same strategies are not used in every case of performed analyses.

Depending on the various problems and research areas in which the aim is to identify trends relating to the development of a specific phenomenon or technology, a different search and identification strategy is established each time, as well as the use of different tool sets. The key issues here are related to the given research area, the types of documents that the researcher will process further, the applied method of collecting and exploring information (including metadata), and above all, having appropriate

knowledge and skills to use relevant information resources adequately to the given situation.

This is complemented by tools provided by both traditional and digital libraries or archives. These tools provide one (though not the only) primary means of analysing technology trends based on the current literature on a particular research problem and its accompanying metadata. Importantly, this literature is not limited to academic papers (field know-how). Libraries today, especially academic and professional libraries, provide a number of tools that allow access to various types of documents, especially the so-called grey literature.<sup>1</sup>

Discovery systems/multi-search engines enable rapid discovery and user-friendly delivery of resources to which a library (or consortium of libraries) has access. These multi-search engines offer integrated search of resources, including those belonging to the so-called hidden internet (e.g. library catalogues, field-specific databases). The use of selected discovery systems makes it possible to identify the situation in the research, development and innovation (R&D&I) area, especially in the context of trend-setting based on the most recent scientific literature. It also allows a preliminary analysis of the phenomena in the environment and to indicate the variability of this environment over time (in particular over particular years or specific geographical areas). From the point of view of access to information on scientific publishing production, the use of the WorldCat central catalogue also proves to be important. It should be noted, however, that the catalogue itself is not a discovery system. However, due to the indexing of information about the resources of libraries in nearly the entire world (also in Poland), in this case it is also worth noting the possibility of its use during the process of programming the support for R&D&I activity. WorldCat allows narrowing search results using the *Topic* category, which is a kind of systematic arrangement in the form of a long drop-down list of keywords having much in common with the classes of the information and search language of the Library of Congress Classification [21].

In the course of the search and identification of sources, Bradford's Law (after Samuel C. Bradford, a librarian of the Science Museum in London) is applied, according to which as many as 1/3 of the publications concerning a given topic were published in works not belonging to a given field. Publicly available, open and free sources of information (including open data) are considered. In the case of Big Data mining, it is necessary to use an appropriate typology of documents, which will enable proper organization of work related first to searching, organizing and structuring of collected content, and will then facilitate processing, data analysis and visualization of the results of these activities. Data can come from a variety of sources, including unstructured sources—it is currently estimated that over 90% of information is in fact stored in unstructured form [22]. For the purpose of source identification, two typologies of documents are used: the classic, which refers to the traditional division of documents into primary, secondary and

---

<sup>1</sup> Anything that is not controlled by publishers, e.g., government, academic, industrial, business materials in print or electronic form, falls into this group of information resources. The latter group of documents includes scientific, technical, economic, social and other reports from public and private institutions, conference materials that are unpublished and not available through commercial publishing or bookshop chains, technical standards and recommendations, unpublished translations, articles published in short-run periodicals that are made available free of charge or of local circulation, certain official documents, technical, promotional and advertising documentation, and documents in electronic form.

derivative, and the typology strictly related to digital resources. Both typologies provide a starting point for collecting data from various sources (for instance—.edu, .gov, .com or .org).

All content about systems, processes and events in the economy, but from different sources, is considered. They, therefore, include both standard and specialized (e.g. scientific) search engines, data from public institutions (e.g. government, local government, European Union), commercial organizations, bibliographic (abstract) and full-text databases, patent databases, social networking sites, blogs, microblogs and discussion forums, RSS feeds. This group also includes data and documents presented in an open way, library catalogues and discovery systems, thematic and business catalogues, databases about companies, content included in information portals.

Data are collected from a wide variety of data sources, mostly available via the Internet, in particular, Open Access: data from public institutions, government, commercial organizations, social media, patent databases, grey literature, information portals, media, commercial organizations, open data and others. In the majority, unstructured data are collected.

Access to publications is accomplished in two ways. The first, the so-called 'gold route,' involves publishing in peer-reviewed journals. The second, so-called 'green way' is for researchers to make their publications, including non-peer-reviewed publications, available in open digital repositories and libraries (self-archiving). For the past decade or so, both the number of journals and repositories have been growing rapidly. In the case of journals, these are obviously new periodicals, but also existing ones that have changed their publishing model to Open Access. Suber observes that an increasing number of these titles "are gaining a reputation as professional, prestigious periodicals of great scientific importance [...] There is also a growing awareness that supporting OA journals is an effective way to support science, researchers, scientific institutions, and the review process. More and more publishers of journals that have become open access journals are noticing that their citation rates have increased, as well as the number of papers retrieved" [23]. Open access journals, digital libraries and repositories are being introduced by the best universities in the world (e.g. Harvard, MIT or Princeton), which is directly related to the promotion of science, but also to strengthening its role in the economic development of countries.

Given the crucial importance of the reliability of the source materials used in the system for identifying technology trends, empirical research was conducted to verify the reliability of open data. To this end, a comparison was made between the results of analyses stemming from the documents from publicly available (open) sources and those stemming from the analyses of licensed (closed) sources.

In order to feed the Analytical Warehouse with data related to innovation and technology, resources were identified and entered into the database. This facilitated the start of the process of automatic learning and assignment of data sources concerning a given field to specific categories: business, science, administration, patents, tenders, media and communication, and Web 2.0.

Big Data analysis is based on data in the form of plain text encoded in UTF-8 standard. In this form, they are stored in the repository. Data must be converted to this form if it exists in another format. Mechanisms have been developed for conversion from PDF

(most scientific publications are downloaded in this format), DOC, DOCX and ODT (a small number of scientific publications), HTML (e.g. patents on the Google Patents platform) and RSS (website news feeds). In the case of PDF format, a small part of the source data requires an additional operation to extract text from the image using OCR technology. In the following, the file formats mentioned above are referred to as non-text files. Input information is provided primarily in the form of URL links directing to web pages containing data (e.g., search results from Google Patents), web pages containing links to non-text files with data (e.g., scientific journals, results of DOAJ repository searches), direct links to non-text files, RSS feed addresses.

For the most part, neural networks are used to identify the values of metadata that constitute the logical structure of the analytical warehouse, e.g. for source classification, document classification into the given field, or publication language recognition. A different approach was used for each of the above cases. Document classification into the given field—since in each task a different field will be given, the model must be built on the basis of provided descriptions. In this case, we will rely on the corpus and the neural network provided by the model builder Google (the BERT model—the latest model developed for natural language processing tasks). For publication language recognition, a model developed by Google was used. It is based on neural networks. In this project, we will not build a neural network model; we will use the networks provided with the model. An automatic data classifier was created as part of the system. Supervised learning algorithms and natural language processing algorithms were used for this purpose. After data preparation, a division was made between a learning set and a test set. The learning set contains the data on which the machine learning algorithms train the model, and the test set is the set left to confirm the classification results of the learned model. The test set is not used during training because as a rule it should be invisible for the model so that the test results are reliable. The processed input of the learning set was fed to the deep network, which learned its internal dependency representations and created a classification model. The model thus created was used to predict labels for a test set that underwent an identical transformation with the fastText algorithm. The prediction result for each record of the test set is a probability vector at the position corresponding to the label position in the list.

One of the most important considerations when using tools that automate the process of downloading data from the Internet is defining an appropriate value for the 'User Agent' parameter. This parameter is a kind of 'business card' of a web browser. Thanks to this value, the server learns which browser the request comes from. Some servers do not allow an empty 'User Agent' value or vary the response content depending on the value of this parameter. Therefore, care must be taken to ensure that each of the tools used for data acquisition allows this to be defined. When retrieving data for Big Data analysis, the 'User Agent' parameter is set to identify the current version of the Chrome browser [e.g., 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/60.0.3112.113 Safari/537.36'].

The database constituting the Analytical Warehouse accepts sources in the form of RSS feeds, which are then regularly checked for new messages, which poses a problem when feeding the database with text document files. To solve this problem, proprietary software was created that creates 'artificial' RSS feeds, also referred to as quasi-RSS,

from downloaded documents. The content of an RSS feed includes a list of ‘messages,’ each with a date associated with the file and a link leading to the text form of the document located on the server. An RSS feed prepared in this way can be added to a database, where it will be processed in the same way as other standard RSS feeds.

The process of creating quasi-RSS files is performed by a custom tool written in Perl called `make_quasi_rss`. The input is a tree structure of directories containing non-text files, created by the `doaj_fetch` or `doaj_fetch_batch` tool, or any directory containing non-text files at any nesting level. The program performs data extraction, converts the non-text file to text, saves the resulting text in a UTF-8 encoded text file under a unique name obtained from calculating the sha256 checksum of the input non-text file. On the basis of the obtained text files, an XML file compliant with the RSS specification is created containing links to the created text files. The program saves information in the database such as the original file name, the sha256 checksum, the date from the metadata, information about whether the file requires an OCR process and whether the conversion proceeded without error.

The database used to conduct the research includes sources from many different services and in different formats. Therefore, before starting computations for each study, the database must be checked for completeness of the sources it contains dedicated to the study that were previously added. Depending on the type of source, the verification method may vary. A quantitative method compares the number of sources downloaded by proprietary software to the number of sources in the database. In case of a difference of more than 10%, the process of downloading and adding sources to the database is repeated. This method consists in comparing individual items from the list of sources prepared for the study with the sources present in the database. If any source is missing, it is added again to the database until the match between the source list and the database is 100%.

The analytical tools applied in the system for identifying technology trends are used in the following sequence of operations: tokenisation; quantitative analysis; TF-IDF statistics; bigram analysis; correlation analysis; cluster analysis.

The search and identification of technology trends take place on the basis of source materials belonging to the field indicated. The classification of industries is intended to define the area to which the document applies. Industry classification is made at the document level. Not every document can be clearly classified into the right category. Documents that cannot be assigned to a category with the appropriate confidence level cannot be mapped. The confidence level should be determined by the machine learning model based on the statistic denoting the similarity of text classification. The source documents should be automatically classified into topic groups using machine learning methods. The machine learning model performs the transformation of the category descriptions into numerical form. The same operation is applied to the texts of publications and the measure of similarity is determined. A sufficiently high similarity between the class and the text allows text classification. Due to the large number of documents to be assigned, the classical supervised method of labelling a sample of articles for each code (class) and using a machine learning algorithm would require a huge amount of effort to collect and manually assign the learning material, which would not at all guarantee a good classification result. Therefore, an unsupervised method based on the ready-made BERT

language model distributed in open-source mode by Google was chosen for the task. BERT is the latest model developed for natural language processing tasks, which is based on the architecture of deep neural networks learned on massive text sets. BERT stores representations of words in over 100 languages as tensors in a multidimensional space and allows automatic selection of the appropriate representation depending on the context of the word. Performance on standard language tasks places the BERT model at the top of the group of ready-made models. The advantage of BERT over other models is its ability to automatically create tensors for whole sentences and texts in different languages. Thus, texts of different lengths can be compared for similarity.

In order to assign a PKD code (Polish Classification of Activity) to texts, each code description was first transformed into a 768-dimensional tensor using the BERT model. The resulting tensors were saved in HDF format for quick access in the next step. The code transformation is required only once, and the finished tensors created on both languages can be used automatically in new iterations on new data. After converting the codes, each English text was also converted to tensors using the BERT model with the same parameters as the PKD codes. After converting the texts, the tensors were used to calculate a similarity measure between the text and the PKD codes. The distance in Euclidean space, which is used in cases of this type, was used as a measure. For each text tensor, a similarity measure was created with the PKD code. The most similar PKD tensor was selected, and its code was finally assigned to the text. In addition to the code, a distance measure in normalized form (range 0–1) was also assigned to each text. The entire operation was performed automatically with the FAISS library, which was made available on an open-source basis by Facebook. FAISS is used for express computation of vector similarity on a large amount of data, which allows very efficient computation compared to classical methods for computing distances in multidimensional space. Analogically, for the assignment of PKD codes in English, tests for Polish were carried out using the BERT method. After analysing the results, it was found that the algorithm did not perform very well due to the accident of the Polish language. The verbal forms that occurred in the texts did not translate into the forms found in the PKD codes. One could try to solve this problem by reducing all words to their basic form (lemmatization); however, there is no open model that would be able to perform this operation in a satisfactory time on the Polish language. For this reason, it was decided to use another unsupervised method, namely the fuzzy matching method.

In order to obtain the greatest possible substantive value of the results of research conducted in the system for identifying technology trends, it was necessary to launch a procedure classifying materials for a given field according to the criteria of category. Source classification is done to recognize where a document comes from to be able to use the specifics of the source. The problem is to seek differentiated assessments of identified technology trends. For example, the categories business, tenders, patents are a form of indication of the intensity of technology consumption, while science assesses the dynamics and prediction of technology change. Documents from social networking sites may be interpreted differently than scientific or patent publications for research on technology trends. In addition, it is possible to identify differences in the popularity of technology between information from the business and academic worlds based on data from different categories.



In order to create a mechanism for automatic publication categorisation, the learning set was classified. Each record was manually assigned by experts to one of seven thematic groups.

The R scripting language was used to create an automatic data classifier using supervised learning and natural language processing algorithms. The input data was developed in an Excel spreadsheet containing records divided into columns, and the result of the script is a finished model and measures of its effectiveness.

Publications that were not ranked by experts were ranked using a machine learning model. Supervised learning and natural language processing algorithms were used to create an automatic data classifier. The input data were unclassified records representing data sources. The result was a finished model and measures of its effectiveness.

Model verification consisted of comparing the category prediction made by the model with the classification made by the experts. For this purpose, a test set was used that the learning model was not familiar with. A confusion matrix was created to compare the categories given by the model with the human-made categorisation. The whole model accuracy—the number of correctly performed categorisations for all test cases is 0.84%. The trained machine learning model was used to map the remaining sources, which were not assigned by experts.

## Discussion and evaluation

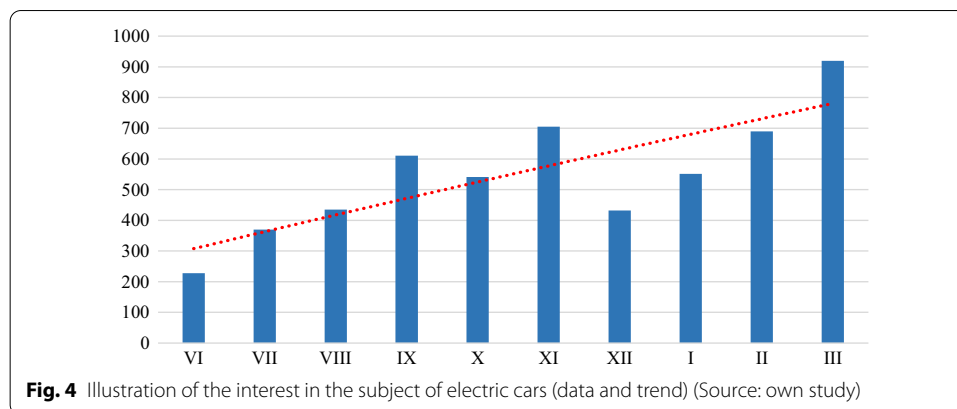
In collecting, storing and analysing data, the presented system uses elements of artificial intelligence, in particular semantic Big Data analysis using advanced natural language processing methods. The applied data refinement method includes cleaning, lemmatisation and refinement of data collected in identified and clustered big information resources.

The purpose of the tool is not only to identify trends relating to the development of a particular phenomenon or technology, but also to precisely define the search strategy itself and to identify and use a particular set of analytical tools. To this end, the tools developed are based on advanced artificial intelligence solutions. In the course of testing and using the tool presented here, several key issues were identified and diagnosed that may hinder wider use of the presented solution. Among other things, the lack of, or insufficient knowledge and ability to use relevant information resources appropriately for a given situation was identified as a key barrier.

All content about systems, processes and events in the economy, but from different sources, is considered. Familiarity with the processes themselves can also be a significant barrier to effective use of the tool by employees of public institutions.

In addition, the procedure that classifies materials for a given discipline according to categorical criteria can be a significant limitation—changing (external) categorisation criteria and catalogues, which is the result of changes in science policy, can greatly limit the formal applicability of the results of the analyses conducted.

However, potential problems or risks associated with the management of the process of analysis and use of information from the tool do not change the fact that the solution has great potential for implementation. This solution allows a more effective science policy to be pursued by the institutions called upon to do so.



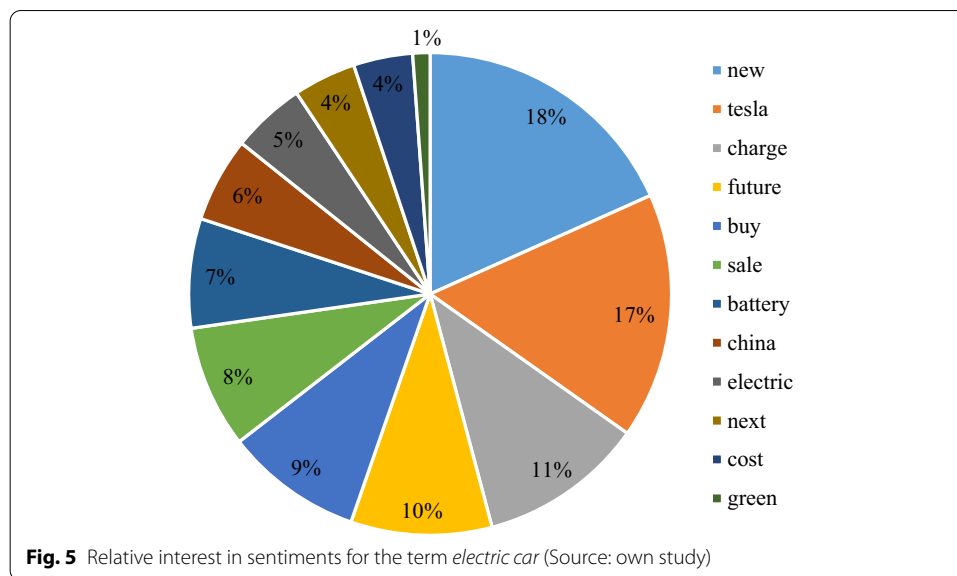
The model verification showed that, in the comparison of the predictions made by the described tools with the classifications made by experts, the former are more accurate and precise. Moreover, the results of the analyses performed by the presented model are not influenced by distorting factors—fads, trends, political pressures, or processes with an unidentified, non-substantive background. It should be emphasized that the accuracy of the whole model is 0.84. Over time, this relevance increases as the applied machine learning model using artificial intelligence algorithms learns during successive iterations of the analytical process.

As part of tool testing, a test identification of innovation processes and monitoring of selected technology trends was carried out. Test results suggest that the tool presented here will contribute to a high degree to a more efficient allocation of resources within the framework of support instruments used by public institutions.

A good example of the functionality of the system and the possibilities of its use in planning public policies is the analysis of the *electric car* technology trend for planning R&D support tools in this area. For the purposes of this study and to analyse the technology trend in *electric cars*, 42 sources were identified, from which 2 TB of data were collected for the period of June 2017–March 2018. Unstructured text sets were processed and analyses based on sentiment analysis were performed. We identified a systematic increase in interest in *electric cars* during a particular period (trend—the dotted line, Fig. 4),

The distinguishing sentiments were as follows: *new*, *Tesla*, *charge*, *future*, *buy*, *sale*, *battery*, *china*, *electric*, *next*, *cost*, *green* (Fig. 5). This reflects the relative interest in the distinguished sentiments, i.e. words-terms that occurred most often in the vicinity of the term *electric car*.

Interest in the issue of *electric cars* grew invariably in the analysed period. Other sentiments than those shown in the figure, have a lower frequency of occurrence. On the other hand, the term “*new*” is distinguished by the intensity of occurrence and dynamics of growth. “*Future*” is a similarly growing sentiment. These increases are accompanied by the growing frequency of the “*Tesla*” sentiment, which is synonymous with novelty (in the context of *electric cars*) and in fact is a brand with an image permanently associated with electromobility. It seems that this company is currently the most important reference point for assessing the phenomenon of *electric cars*. “*Battery charging*” sentiment turns out to be a practically very rudimentary issue, which is stable with a downward



trend. The decline of “China” as a sentiment of *electric car* is a distinguished issue. This may indicate the country’s declining importance for the development of electric cars. The decreasing interest in the costs related to *electric cars* can be associated with the emerging capitalization of the already existing technical and market solutions for *electric cars* and, at the same time, the growing influence of environmental protection issues on the development of electromobility. The association of *electric cars* with such sentiments as “green” is invariably and relatively fast-growing. The results of our study show the usefulness of the described procedure and IT system.

## Conclusions

Modern methods of identification and diagnosis of technology trends, presented in the article, allow better adjustment of the support offered for R&D activity, but also make it possible to apply and improve management methods in public institutions. Moreover, the presented IT solution allows obtaining highly reliable data. The accuracy of the entire model is 0.84.

The actual technological and scientific potential of beneficiaries of R&D support programmes can be determined with methods that go far beyond the existing research and analytical tools used by public institutions.

The system for identifying technology trends presented in this article is one proposal for a solution which enables more effective management of public funds directed to the development of innovative activity. The presented solution makes it possible for public funds at the disposal of support institutions to be spent in a way which will be better adjusted not only to the needs of scientific institutions and enterprises carrying out R&D activity, but also to the scientific and technical potential of these organizations.

The tools presented make it possible not only to distinguish strategic sources of information for identifying technology trends, but above all to automatically collect and process unstructured data. The tools allow quantitative analysis of this data, prediction in defined areas and visualization of the results of analysis. The solution we use can be “fed”

with data from different sources. We presented the possibilities of using online data from the Internet. In the near future, we plan to calibrate the IT system driving this solution for the use of data from scientific libraries and knowledge repositories in scientific institutions.

Tools similar to those described in this article are used in various projects, research activities and managerial processes, not only in organisational routines, but for reconfiguring the organisation's resource base as well. Despite the fact that similar tools are currently used in the world for the above-mentioned tasks, only our solution was a fully functional tool to support management processes in public administration as early as 2018. All functionalities that support institutions to conduct more accurate process of designing public instruments supporting research and development activities were already in action then. Compared to similar solutions, our tools were one of the first among modern management supporting IT facilities, and the first to support the achievement of public policy goals.

The case studies presented in this article show the potential of the tools used, including the possibility of their application in research on both technological topics and social research related to policy making.

Our team is focused on the further improvement and development of this tool. We plan to test it in research on new fields related to technology, including challenges resulting from combining technologies from various fields (e.g. biomedicine and materials technology). The obtained results suggested we plan the application of our solution in the fields of social sciences, including economic sciences and management, as well. From that point of view, our tools will be tested in the near future in a research project focused on the labor demand and research on disinformation and fake news, which in the current geopolitical situation has grown to be of considerable importance.

#### **Abbreviations**

AI: Artificial intelligence; GUS: Statistics Poland (Polish national statistical office); IT: Information technologies; PKD code: Polish Classification of Activity code; R&D: Research and development; R&D&I: Research, development and innovation; RINGf system: Information Refining System.

#### **Acknowledgements**

The authors thank the National Center for Research and Development for funding the project and the Faculty of Political Science and International Studies for supporting its implementation.

#### **Author contributions**

All authors' contributions are equal. All authors read and approved the final manuscript.

#### **Funding**

National Center for Research and Development (Cybersecident/489281/IV/NCBR/2021), University of Warsaw.

#### **Availability of data and materials**

In the text.

#### **Declarations**

##### **Ethics approval and consent to participate**

Not applicable.

##### **Competing interests**

Not applicable.

#### **Author details**

<sup>1</sup>The Faculty of Political Science and International Studies, Krakowskie Przedmieście 26/28, 00-928 Warsaw, Poland. <sup>2</sup>The Faculty of Journalism Information and Book Studies, Krakowskie Przedmieście 26/28, 00-928 Warsaw, Poland. <sup>3</sup>Institute of Economics of the Polish Academy of Sciences, Staszic Palace, 72 Nowy Świat St., Room. 266, 00-330 Warsaw, Poland.

Received: 2 January 2022 Accepted: 6 April 2022

Published online: 28 April 2022

## References

- Schumpeter JA. Capitalism, socialism and democracy. Crows Nest: George Allen & Unwin (Publishers) Ltd; 2003. (1976, Edition published in the Taylor & Francise-Library).
- Żołnierski A. Nieformalne źródła informacji w działalności gospodarczej. In: Cetera W, Kowalik K, editors. Logistyka i administrowanie w mediach. Instytut Dziennikarstwa Uniwersytetu Warszawskiego: Warsaw; 2015.
- Ohlhorst FJ. Big data analytics: turning big data into big money. Hoboken: Wiley; 2015.
- Chai S, Shih W. Why big data isn't enough. MIT Sloan Manag Rev. 2017;58(2):57.
- Rao R. From unstructured data to actionable intelligence. IT Prof. 2003;5(6):29–35. <https://doi.org/10.1109/MITP.2003.1254966>.
- Je S-M, Ko H, Huh J-H. Accurate demand forecasting: a flexible and balanced electric power production big data virtualization based on photovoltaic power plant. Energies. 2021;14(21):6915. <https://doi.org/10.3390/en14216915>.
- Tran DT, Huh JH. Building a model to exploit association rules and analyze purchasing behavior based on rough set theory. J Supercomput. 2022. <https://doi.org/10.1007/s11227-021-04275-5>.
- Richards G. Big data and analytics applications in government: current practices and future opportunities. Boca Raton: CRC Press; 2018.
- Frishammar J, Richtnér A. Editorial. Int J Technol Intell Plan. 2008;4(3). [International Journal of Technology Intelligence and Planning \(IJTIP\) Inderscience Publishers—linking academia, business and industry through research.](https://doi.org/10.1007/s11227-021-04275-5)
- Hayashi AM. Thriving in a big data world. MIT Sloan Manag Rev. 2014;55(2):35.
- Ben-Hur S, Jaworski B, Gray D. Aligning corporate learning with strategy. MIT Sloan Manag Rev. 2015;57(1):53.
- Jain S, Builteir de A, Fallon E. An extensible parsing pipeline for unstructured data processing. In: Conference: 2021 23rd international conference on advanced communication technology (ICACT), February 2021. <https://doi.org/10.23919/ICACT51234.2021.9370654>.
- Balabin H, Hoyt CT, Birkenbihl C, Gyori BM, Bachman J, Kodamullil AT, et al. STonKGs: a sophisticated transformer trained on biomedical text and knowledge graphs. Bioinformatics. 2022;38(6):1648–56. <https://doi.org/10.1093/bioinformatics/btac001>.
- Pagad NS, Almuzaini KK, Maheshwari M, Gangodkar D, Shukla P, et al. Clinical text data categorization and feature extraction using medical-fissure algorithm and Neg-Seq algorithm. Comput Intell Neurosci. 2022. <https://doi.org/10.1155/2022/5759521>.
- Pal Nandi B, Jain A, Tayal DK, Narang PA. High performing sentiment analysis based on fast Fourier transform over temporal intuitionistic fuzzy value. Soft Comput Fusion Found Methodol Appl. 2022;26(6):3059–73. <https://doi.org/10.1007/s00500-021-06444>.
- Azad P, Navimipour NJ, Rahmani AM, et al. The role of structured and unstructured data managing mechanisms in the Internet of things. Clust Comput. 2020;23:1185–98. <https://doi.org/10.1007/s10586-019-02986-2>.
- 2021 analytics research review: global market insights for video, text, and marketing analytics with projections to 2026, M2PressWIRE, 2022 Mar 9. <https://search.ebscohost.com/login.aspx?direct=true&db=nfh&AN=16PU1135620601&lang=pl&site=ehost-live>.
- Habermas J. On the pragmatics of communication. Cambridge: MIT Press; 1998.
- De Moor K, Berte K, De Marez L, Joseph W, Deryckere T, Martens L. User-driven innovation? Challenges of user involvement in future technology analysis. Sci Public Policy. 2010;37(1):51–61.
- Stephens-Davidowitz S. Everybody lies: big data, new data, and what the internet can tell us about who we really are. New York: HarperCollins; 2017.
- Hollender H. Deskryptory i przyszłość opracowania rzeczowego w bibliotekach. Rocznik Biblioteki Narodowej. 2015;46:406.
- Gang-Hoon K, Trimi S, Ji-Hyong C. Big-data applications in the government sector. Commun ACM. 2014;57(3):78.
- Poynder R. Suber – przywódca ruchu bez lidera. Biuletyn EBIB; 2011. nr 7, p. 3–4. [www.ebib.pl/images/stories/numery/125/125\\_suber.pdf](http://www.ebib.pl/images/stories/numery/125/125_suber.pdf).

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.