The European Proceedings of
Social and Behavioural Sciences
EpSBS

www.europeanproceedings.com
e-ISSN: 2357-1330

DOI: 10.15405/epsbs.2020.10.03.184

ICEST 2020

International Conference on Economic and Social Trends for Sustainability of Modern Society

PROBLEMS OF OPEN DATA SOURCES ANALYSIS FOR SOCIO-ECONOMIC AND MEDICAL RESEARCH

Anna M. Kashirina (a)*, Anatoly V. Kravchenko (b)
*Corresponding author

(a) Novosibirsk State Technical University, K. Marx Ave., 20, 630073, Novosibirsk, Russia, kashirina@corp.nstu.ru
(b) Novosibirsk State Technical University, K. Marx Ave., 20, 630073, Novosibirsk, Russia, a.kravchenko@corp.nstu.ru.

Abstract

Socio-economic studies cover a wide range of issues related to expanded reproduction, gradual structural and qualitative changes in the economy, production forces, growth and development factors, science, education, culture, quality and standard of living of society, human capital. The goals of socio-economic development of the regions may be to improve the quality of healthcare, education, increase income, etc. According to the development goals, criteria and indicators are determined on the basis of which these criteria will be measured. Their measurement requires large amounts of statistical data. Of particular relevance in the present time is the process of obtaining reliable and timely data on the level of disease in the world and Russia in a pandemic. Most of the social and economic resources of the Russian Federation are formed by government departments and commercial organizations during marketing, analytical and consulting studies and by teams of specialists, including in the field of healthcare. The format of such data should enable further automatic processing and visualization. Statements published on the internet resources do not always meet the requirements of stakeholders. In the course of the study, databases on socio-economic studies, in particular health care (open access, available for download on the Internet) were studied, data presentation formats and the possibility of their further processing were analyzed. An analysis of 10 public data sources for health information and research results on the incidence in the Novosibirsk region has been fulfilled.

Keywords: Open data, data analysis, socio-economic research, healthcare.

https://doi.org/10.15405/epsbs.2020.10.03.184
Corresponding Author: Anna M. Kashirina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

1. Introduction

According to the Strategy for the Development of the Information Society in the Russian Federation for 2017 - 2030, one of the main directions of the development of Russia for the next decade will be to ensure the development of the economy and social sphere (Kashirina, 2019). To attain its goals, it is necessary to receive and process large amounts of social and economic information in a timely and efficient manner. In the same time, the new requirements for socio-economic reporting are the accumulation, systematization, unification and the possibility of access to information that was previously disparate in nature. An extremely important aspect is the possibility of access to an information resource that allows us to receive information about socio-economic activity. The directions of using social reporting are very diverse, social reporting is used for (Alekseeva et al, 2019): analysis of firm resilience (Koç & Durmaz, 2015), identifying and analyzing the most significant stakeholders (Şener et al., 2016), confirming the relationship between social reporting costs and improving financial performance of companies (Lys et al., 2015), confirming the impact of reporting corporate social responsibility on the financial performance of firms (Abukil, 2016), analysis of the impact of labor indicators and human rights on firm value (Mulya, 2018), and even the study of the relationship between the gender of the head of the organization and the level of disclosure (Mohd-Said, 2018).

2. Problem Statement

In the early 2000s, scientists at the Research Computer Center of Moscow State University and the ANO Information Research Center conducted a study of databases on socio-economic statistics of the Russian Federation with the possibility of Internet access. The main way of obtaining such information was the University Information System Russia, containing data on socio-economic statistics. About 20 years have passed, which are characterized by an increase in demand for information, an increase in the number of information services. Information research has become an integral part of business life. We will consider what has changed during this time, what opportunities for access, analysis, processing and visualization of socio-economic data have appeared.

3. Research Questions

Socio-economic studies cover the level and quality of life, economic activity, accessibility and quality of social services, education, healthcare, etc. The object of socio-economic research is the population, the activities of enterprises, regional and municipal development, the activities of state and society organizations. Environmental, social, and managerial challenges are becoming a powerful means of gaining a competitive edge in the global market (Bîltac, 2019).

Social and economic research statistics have always been the basis for research projects. Most of the statistical resources of the Russian Federation are generated by government departments and commercial organizations in the course of marketing research, scientific researchers at universities, analytical and consulting centers, during surveys and questioning of the population. With the growth of data volumes, the need for online databases became obvious, where statistics collected by specialists would be accumulated, work was done on their storage, processing, updating, verification, description, translation into a single

https://doi.org/10.15405/epsbs.2020.10.03.184
Corresponding Author: Anna M. Kashirina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

format, the possibility of uploading for further processing, analysis and visualizations for a more visual presentation (Bogomolova et al., 2003).

Socio-economic development is a complex phenomenon, the analysis and forecasting of which requires a large amount of data describing the state of not only the socio-economic system itself, but also related areas. General requirements for the source data: quality, adequacy, reliability, accessibility of the methodology for the formation of indicators, the necessary accuracy, comparability, consistency, practical usefulness, consistency, scientificness, completeness, comprehensiveness, sufficiency, the effectiveness of the development of statistical data, timeliness, interpretability, appropriateness, openness and accessibility, ease of presentation of statistical data, confidentiality (Marshova, 2017).

Statements found on the internet resources, do not always meet the requirements of the stakeholders. There is a problem related to the timeliness of the information contained in the financial statements, the weak performance of site updates, refresh rate of the information provided, the weak performance feedback (Kamalluarifin, 2016). Information technology, in particular robustness, is a tool for maintaining timely and effective communication with stakeholders (Bosetti, 2018).

The Federal State Statistics Service defines open data as information on the activities of state bodies and local authorities posted on the Internet in the form of data arrays in a format that ensures their automatic processing for reuse without prior change by the person. and on the terms of its free use. From the definition it is clear that the format of this data should allow for further automated processing. Thus, the objective of this study was to find a database of open government data on socio-economic research, which, firstly, are available for download on the Internet, and secondly, a format which allows you to effortlessly carry out further processing. The first steps towards the formation of accessible arrays of open data were made back in the 50s of the last century (Radchenko, 2013). The first steps towards the formation of accessible arrays of open data were made back in the 50s of the last century (Radchenko, 2013). Open government data appeared in the Russian Federation in 2012. In 2013, the countries participating in the G8 summit adopted the Charter of Open Data, which provides for public disclosure of information by government agencies on the Internet. In 2006-2015, a number of laws and regulations were adopted in the Russian Federation in support of the government exchange of information. In 2014, the data.gov.ru portal was launched, currently more than 23755 data sets have been published there (Koznov et al., 2016).

In Russia, many government agencies, private and uncommercial organizations publish significant amounts of open data for public use. In particular, multi-published collections containing official statistical information that reflects the phenomena and processes that have taken place in the economic and social life of the Russian Federation, gives a complete picture of changes in the country during different periods of Federal districts, Republics, areas, regions. But, as a rule, such collections contain text files that cannot be processed without additional transformations, which sometimes require a lot of time. Usually this is a doc- or pdf- format that does not allow you to immediately convert them to xls- tables, on the basis of which analysis is usually carried out, forecasts, charts are built, infographics or other visualizations are done in the Excel spreadsheet editor. And even if the service offers the ability to upload data in xls-format, they need additional processing before they are used for analysis. Such tables contain combined cells and header lines, text number formats, they cannot be filtered due to the presence of subheadings or notes in the table, data for different periods are on different sheets, etc.

https://doi.org/10.15405/epsbs.2020.10.03.184
Corresponding Author: Anna M. Kashirina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

4. Purpose of the Study

The methods for presenting information on such sites need improvement. There is no possibility of filtering data for detailed sampling: the choice of viewing and downloading periods, the nomenclature (products, types of economic activity, etc.); there are no long dynamic series. Thus, users have to select information from different statistical collections and tables, convert them into tables on their own for further analysis and processing, which is time-consuming (Marshova, 2017).

5. Research Methods

Data analysis technology to prepare for subsequent publication consists of the following steps:

- Search and selection of data from open sources.
- Data cleaning. Data cleaning tasks include error correction: inconsistency of information, omissions in data, anomalous values, data entry errors.
- Visualization. Proper data visualization is one of the key factors for the successful result of the entire analysis. Socio-economic information, depending on its type, needs various forms of presentation (Kashirina, 2016).

In the framework of this study, the authors conducted an analysis of sources, open data, the result is presented in Table 01.

Table 01. Comparative analysis of open data on socio-economic research

Portal name	Dataset groups	Number of indicators	Period, years
Electronic region	9	77	2004-2014
ICSI	12	50	1999-2018
Indicators of municipalities	23	No data*	1999-2019
Official statistics of Rosstat	23	No data*	1970-2018
The World Bank	20	2029	1960-2019
UIS Russia	23	3056	1989-2019
EMISS	8	6725	2002-2020
Multistat	22	761	1970-2013
Statistics Showcase	30	No data*	1990-2020
Open data of Russia	16	23755	No data*

*Note: “No data” implies that the amount of data presented is large enough, but there is no information on the quantity on the site, and calculation is difficult

Let us consider each of the sources in detail.

1. The “Electronic Region” portal, developed by specialists of the Institute for the Development of the Information Society, contains a data analysis section to assess the readiness of the Russian regions for the information society (http://eregion.ru/analiz-dannykh). In particular, it contains data on such factors of the development of the information society as human capital and the economic environment. A total of 77 variables are presented on the portal, grouped into 9 sub-indexes. The data is shown in the form of tables

https://doi.org/10.15405/epsbs.2020.10.03.184
Corresponding Author: Anna M. Kashirina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

and graphs, there is a filter for selecting regions and a period, text manuals contain data for 2004-2014, and you can work on the website in tabular form only with data for the period 2011-2014. The pdf- and xls-table data upload format.

2. The main socio-economic indicators of the Institute for Complex Strategic Studies, ICSS (https://icss.ru/macro2) data on 50 indicators (GDP, prices, social indicators, etc.). From 12 groups in the form of tables and graphs for 1999-2018 year. Download format is jpg- and xls-. Graphs are built on the site, when hovering over - a description of the data point, there are filters (period selection, relative / absolute values).

3. The database of indicators of municipalities of Rosstat contains information on 23 indicators for 2006-2018 year. First you need to choose the territory, period and indicator of interest from 33. Additional filters allow you to choose the municipal district, type of settlement, types of economic activity and period. The result of the request is displayed on the screen; if necessary, it can be saved in xls and csv format. There is no visualization on the site.

4. Statistics collection "Regions of Russia. Socio-economic indicators" Rosstat (https://gks.ru/folder/210/document/13204). The collection is available for download in doc, pdf. The annex to the collection is uploaded in the archive, contains 24 xls-files, which presents data on 22 sections (Population, Living Standards, Health Care, etc.) for 1970-2018. In the Statistics section, there are several options for the finished infographic.

5. The World Bank statistics on the Russian Federation (https://data.worldbank.org/country/russian-federation) contains 2029 indicators for 20 groups. Data for 1960-2019 presented in different languages (no Russian) in the form of small graphs on the main indicators with a description of the data point on hover, are downloaded in the form of tables csv, xml, excel. The site has a database of key development indicators in the form of a table, chart and map with filters by year, indicator and country.

6. University Information System Russia, the developer of the Research Computer Center of Moscow State University annually publishes the collection "Regions of Russia. Socio-economic indicators" (https://uisrussia.msu.ru/stat/). The portal contains 23 data sets with 3056 indicators for the period 1989-2019, the result is displayed in a table, and is downloaded in xls- and htm- format.

7. The unified interdepartmental information and statistical system EMISS (https://www.fedstat.ru/) of Rosstat contains 8 topics, 6,725 indicators for 65 departments. The site has a search bar and the possibility of an advanced search by department, subject, period, state program. Data for the period 2002-2020 unloaded in the form of a table, downloaded in xls, xml format (with the ability to select data). On the site you can present the data in the form of a graph or a histogram, when you hover over the cursor, the data signatures appear, there is a calculation procedure, a department, a characteristic and the ability to ask an expert a question. Data can be selected by sinks and columns, filtered by constituent entities of the Russian Federation, hidden, transposed, expanded.

8. The multi-functional statistical portal Multistat Rosstat (http://www.multistat.ru/) provides paid access to the statistical database "Economics of Russian Cities", which consists of 761 indicators for 1970-2013.

9. The statistics showcase (https://showdata.gks.ru) of the Rosstat contains 30 sections with subsections. A report is built on the site in the form of a table and a graph, there are filters by period, region,

https://doi.org/10.15405/epsbs.2020.10.03.184
Corresponding Author: Anna M. Kashirina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

transpose, settings using the constructor, data is downloaded in xls- and csv- format. Plotting provides a choice of 11 types of diagrams.

10. Open data of Russia (https://data.gov.ru) - a single access point to all open state data of the Russian Federation - contains 23755 indicators data in 16 categories, allows you to upload data in CSV- and XML- format.

The analysis showed that there is no one database that meets all the criteria, each of the considered databases has one or more drawbacks that do not allow you to immediately use the data for analysis and visualization. The best results in terms of period and the data presented showed Rosstat's portal. But, as it turned out, the portal has several different platforms on which search results are presented. This is a very difficult job for inexperienced users.

Consider working with sites of statistical databases on the example of searching for information about the disease tuberculosis. As an additional criterion, the NSO region was considered and the possibility of uploading data to Excel tables was evaluated for further processing. In the process faced with several options for the search and, accordingly, with different results, the results are presented in Table 02.

Table 02. Search results on the statistics portal on tuberculosis diseases in the Novosibirsk region.

Site. **Search option**	**Result.**	**Sampling of tuberculosis / NSO data**
Rosstat. Appendix to the statistics book "Regions of Russia"	xls, 2000-2018	-/+
Rosstat. Official statistics	xls, 2000-2018	+/-
Rosstat – Publications	xls, 1991-2018	+/-
The World Bank	csv, xml, xls, 2000-2018	+/-
UIS Russia	xls, 2005-2016	+/-
EMISS	xls, 2005-2016	+/+
A showcase of statistics (Search query)	xls, 1192 -2017	+/+
Novosibirskstat	pdf, doc, 2018-2019	+/+

6. Findings

Let us consider the search results in more detail. The electronic region, ICSI and database indicators of municipalities do not contain health data. The Russian Open Data portal, upon request, in the search bar, provided only data on tuberculosis mortality, which could not be viewed on the website, the data are downloaded in xml format, and special software is used to work with such open data set files. Other sources give the result of the request, but quite different. Rosstat has several portals for working with data. The database is one, but the search options, the ability to obtain data and, accordingly, the result of obtaining information on request is different. Appendix to the statistics collection “Regions of Russia. Socio-economic indicators” is downloaded in a single archive with Excel files, which contain different sections, the contents are in a separate file. To receive a response to the request, the 5th section “Healthcare” was selected, consisting of different sheets with indicators, the table of contents - on a separate sheet (https://gks.ru/folder/210/document/47652). It was not found in the table of contents information upon request.

https://doi.org/10.15405/epsbs.2020.10.03.184
Corresponding Author: Anna M. Kashirina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

A search on the same site, but by directly clicking on the link “Official Statistics” in the “Health Care” section, made it possible to download separately the table “The incidence of the population by main disease classes” (https://www.gks.ru/folder/13721). The results were also negative; there are no information on the incidence of tuberculosis in the table. The next search option on the Rosstat website in the section Publications - Catalog of publications - Statistical publications - Appendix to the yearbook made it possible to download an Excel file with health indicators (https://gks.ru/folder/210). The autofilter in the table showed the presence of three indicators for tuberculosis diseases. There were no samples by region. Thus, this search option showed the largest number of indicators in the healthcare section (133) and a rather large sample size. In total, the annex to the collection contains 24 sections with socio-economic indicators for the Russian Federation from 1991 to 2018 year.

A search on the World Bank portal (a request in English in the quick search bar) displayed the indicator “Tuberculosis incidence” (https://data.worldbank.org/indicator/SH.TBS.INCD?view=chart) in the form of a graph that can be change to a chart or map. Data for different countries are shown below with mini-graphs reflecting the trend. You can choose only the country, the region cannot.

On the UIS Russia portal, the search is difficult because no search form. In the annual publication of Rosstat - Healthcare in Russia - Public health was found the required rate. The sample size is 6 years (2005 - 2016, not all years), the table contains 16 indicators for the incidence of tuberculosis. No sampling by region. There is a separate table grouped by gender, age groups, city / village, children.

The query in the EMISS search bar (https://www.fedstat.ru/indicators/) returned 45 indicators. According to the indicator, a table was constructed with the possibility of filtering (period and region), as well as a graph and a histogram. There are indicators with a filter by gender, age and period. The file is downloaded, including in xls- format, covers 13 years. Thus, the task was completed.

A query in the search line of the Showcase of Statistical Data found several indicators, when you select one of them, the Report opens on the website in the form of a table (or 11 options for charts). There is a filter by year, region, in some indicators there are filters by age, population structure. Information on request was found.

The regional site of Novosibirskstat, which contains 10 health indicators, was also examined. The necessary data were found on it, but only for the period 2018-2019 year, and only in pdf- and doc- format.

In the end, the best result was obtained on the portal Showcases of statistical data. The necessary data were found that are optimal in terms of row length, number of indicators, filtering, downloading and visualization capabilities. But it should be noted that during the initial review of the sources this site was not investigated, because There is no direct link to it on the Rosstat website. The site was found randomly while searching for data. Thus, the only, but significant drawback of it is the inability to get on it, not knowing about its existence.

7. Conclusion

It is not possible to evaluate reliably the economic effect of creating sets of open data. According to Russian experts, direct government revenue from the sale of public sector information is less than the economic effect of directly using public sector information by the business community 10-20 times (Kachan, 2018).

https://doi.org/10.15405/epsbs.2020.10.03.184
Corresponding Author: Anna M. Kashirina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

Today, there is a process of quantitative accumulation of data, and the flow of information that requires deeper understanding and structuring is huge. The level of modern scientific knowledge should be high enough to be able to develop optimal solutions (strategies and programs) and contribute to the full manifestation of the positive consequences of the digitalization process (Kamalluarifin, 2016).

Observing the recommendations in this study and the right balance between modern technology and the required level of information culture of the researcher will give the opportunity to receive high-quality raw data and present them in a format that will increase the level of analytical studies, forecast calculations and taken on the basis of their decisions.

References

Abukil, A. A. (2016). The Effect of Corporate Social Responsibility Reporting on Financial Performance in Libya and Jordan. *International Journal of Economics & Business Administration (IJEBA), 4*(4), 113-122.

Alekseeva, I. V., Fedosova, O. N., & Pryadkina E. A. (2019). Analiz sovremennykh issledovaniy formirovaniya sotsial'noy otchetnosti kommercheskikh organizatsiy v usloviyakh tsifrovoy ekonomik. [Analysis of modern studies of the formation of social reporting of commercial organizations in the digital economy]. *Accounting and Statistics, 4*(56), 10-20. [in Rus]

Bîltac, O. (2019). Company's Social Performance Reporting Based on International Standards: A Comparative Analysis Across Central and Eastern Europe. *European Research Studies Journal, 22*(3), 149-167.

Bogomolova, A. V., Karasev, O. I., Sennov, R. A., & Yudina, T. N. (2003). Bazy dannykh po sotsial'no-ekonomicheskoy statistike Rossiyskoy Federatsii s Internet-dostupom. [Databases on socio-economic statistics of the Russian Federation with Internet access]. *Digital Libraries: Advanced Methods and Technologies, Electronic Collections: Proceedings of the All-Russian Scientific Conference*, 332-338. [in Rus]

Bosetti, L. (2018). Web-based integrated CSR reporting: An empirical analysis. *Symphonya. Emerging Issues in Management, 1*, 18-38.

Kachan, D. A. (2018) Otkrytyye dannyye: analiz tendentsiy [Open Data: An Analysis of Trends] *Digital Transformation, 1*(2), 72–78. [in Rus]

Kamalluarifin, W. G. S. W. (2016). The influence of corporate governance and firm characteristics on the timeliness of corporate internet reporting by top 95 companies in Malaysia. *Procedia Economics and Finance, 35*, 156-165.

Kashirina, A. M. (2016). Vizualizatsiya biznes-informatsii. [Visualization of business information] *Bulletin of scientific conferences, 5-4*(9), 134-135.

Kashirina, A. M. (2019). *Razvitiye informatsionnogo obshchestva: uchebnoe posobiye* [Development of the information society: textbook]. NSTU Publishing House. [in Rus]

Koç, S., & Durmaz, V. (2015). Airport corporate sustainability: an analysis of in dicators reported in the sustainability practices. *Procedia-Social and Behavioral Sciences, 181*, 158-170.

Koznov, D., Andreeva, O., Nikula, U., Maglyas, A., Muromtsev, D., & Radchenko, I. (2016) A Survey of Open Government Data in Russian Federation. *8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Porto, Portugal 9-11 November*.

Lys, T., Naughton, J. P., & Wang, C. (2015). Signaling through corporate accountability reporting. *Journal of Accounting and Economics, 60*(1), 56-72.

Marshova, T. N. (2017). Printsipy formirovaniya statisticheskikh dannykh dlya analiza i prognoza sotsial'no-ekonomicheskogo razvitiya. [Principles of generating statistical data for analysis and forecasting of socio-economic development], *Economic and social-humanitarian studies Publisher: National Research University "Moscow Institute of Electronic Technology" (Moscow), 2*(14), 25-36. [in Rus]

Mohd-Said, R. (2018). Board compositions and social reporting: evidence from Malaysia. *International Journal of Managerial and Financial Accounting, 10*(2), 128-143.

Mulya, H. (2018). The Impact of Sustainability Reports toward the Firm Value. *European Research Studies Journal, 21* (4), 637- 647.

Radchenko, I. A. (2013). Ispol'zovaniye otkrytykh dannykh v nauchnykh issledovaniyakh [The Use of Open Data in Scientific Research]. *Information Society, 1-2*, 93-101. [in Rus]

Şener, I., Varoğlu, A., & Karapolatgil, A. A. (2016). Sustainability Reports Disclosures: Who are the Most Salient Stakeholders*? Procedia-Social and Behavioral Sciences, 235*, 84-92.