





# russian-sme-registry: A command-line tool to make a geocoded dataset of Russian small and medium-sized enterprises from tax service open data

Pavel O. Syomin <sup>1</sup>

<sup>1</sup> Independent researcher, Russia

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Open Journals](#) 

## Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## Summary

The Federal Tax Service of Russia operates an official state registry containing information regarding all Russian small and medium-sized enterprises. The registry's monthly dumps are freely available for download; however, they are large and technically cumbersome. This paper introduces a Python command-line tool that automatically extracts and transforms these dumps to create a handy geocoded tabular panel dataset. This tool aims to facilitate data retrieval, thereby enabling researchers to prioritise their study without the encumbrance of complex data preprocessing.

## Statement of need

A considerable number of academic studies have been conducted in Russia on the subject of small and medium-sized enterprises (SMEs), and their role in the national and regional economies. Since its inception in 2016, the official state Registry of Small and Medium-Sized Enterprises (SME registry) operated by the Federal Tax Service (FTS of Russia) has become the primary source of information about SMEs in the country ([Barinova et al., 2023, p. 26](#)). Researchers widely rely on SME registry data ([Eryomko & Goryunova, 2025](#); [Gumenyuk, 2022](#); [Korchagina, 2025](#); [Razmanova et al., 2022](#); [Zakharova et al., 2023](#)). However, there is evidence to suggest that researchers typically utilize only aggregated statistics from the SME registry website, neglecting to access the underlying disaggregated data, despite the availability of this data at no charge. The issue is likely to lie in the considerable volume and relative complexity of working with raw SME registry data. Only a minority of researchers have developed their own ETL pipelines. These are used for either of the following two methods: (1) with Excel tables that are exported from the Registry's web interface, as described by [Kerimov & Mustafaeva \(2023\)](#); or (2) with downloadable ZIP dumps of the Registry, as outlined by [Kuzora & Natarov \(2022\)](#). The necessity for a readily available, open-source tool for SME Registry data extraction and transformation is becoming increasingly apparent. This tool is intended for two distinct audiences: firstly, researchers who wish to utilise raw SME Registry data but are unable to process it due to a lack of technical expertise; and secondly, researchers who possess the necessary skills but prefer to rely on pre-existing tools rather than developing their own code. The objective of the application is threefold: firstly, to simplify the lives of researchers and save their money; secondly, to assist in the discovery of novel research opportunities enabled by the utilisation of disaggregated SME Registry data; and finally, to mitigate the risk of errors caused by bugs in custom ETL pipelines.

The tool performs the following enhancements to the raw Registry data. First of all, the resulting dataset is significantly smaller, with a reduction from more than 200 Gb of raw data to less than 20 Gb. Secondly, the raw data, which is distributed as a collection of ZIP archives with complex hierarchically structured XML files inside is converted to a flat panel CSV

table. Thirdly, the original data is enriched with three additional variables: annual revenue, expenditure, and average workforce of organisations. Furthermore, the option to filter the SMEs by NACE Rev.2-compatible main activity code directly during the extraction is available. In addition, some rarely used variables are dropped to make the focus on the most important ones. Last but not least, addresses on incorporation are normalised, enriched with municipal codes and geocoded to add geographical coordinates.

It should be noted that SMEs are merely a subset of all organisations and sole traders. In Russia, a business is categorised as an SME if its annual revenue is below 2 billion roubles and its average workforce is fewer than 250 employees. In addition, SMEs must not be under the control of government entities or non-SMEs. Consequently, the dataset created by the application represents a non-representative segment of the entire population of juridical persons and sole traders. Furthermore, the spatial granularity of the data is constrained to the level of settlements, as there is an absence of more detailed address information in the SME Registry. If geocoding down to the level of individual streets and houses or the entire population of Russian organisations or sole traders is required, the Russian Financial Statements Database (RSFD) (Bondarkov et al., 2025) may be used.

## Pipeline

The proposed tool is built on the top of a five-stage data processing pipeline. Each stage of this pipeline is implemented as a separate class. All classes have harmonised high-level APIs. A Python developer is likely to comprehend the code with ease; however, a concise description of each stage is provided below for those who may require it.

## Download

The download stage retrieves web pages containing lists of raw data files, parses them to get file URLs, and downloads these files. The retrieved web pages represent three open datasets provided by the FTS of Russia: [SME Registry](#), [Information about revenue and expenditure of organisations](#), [Information about annual workforce of organisations](#). The download stage has been included in the tool as a helper to automatically obtain more than a hundred of raw data files without the necessity for manual download.

## Extract

The extract stage retrieves valuable information from the raw data files. This information is subsequently stored in CSV files, with each file representing a source file. It can also filter SMEs by main activity code. The purpose of this step is twofold; firstly, to discard irrelevant observations and their attributes, and secondly, to transform the complex hierarchical structure of zipped XML files into a flat table.

## Aggregate

The aggregate stage assembles data from multiple CSV files produced during the extract stage into a single file and removes duplicated observations. Duplication arises from the process of data publication. Each monthly dump comprises the entire SME Registry, and if a particular SME record have not been altered, the rows stemming from disparate dumps will be almost identical, differing solely by the data publication date. Aggregation effectively eliminates duplicates and creates start\_date and end\_date attributes to keep the temporal information.

## Geocode

The geocode stage normalises addresses, resolving any inconsistencies in the nomenclature of regions, districts, cities and settlements. It also adds geographical coordinates and municipal

86 codes of cities or settlements. These enhancements transform the SME Registry data into geo-  
87 data, thereby facilitating its utilisation for a spatial analysis. The geocoding is accomplished  
88 via static lookup tables integrated within the application wheel. The estimated accuracy of  
89 geocoding is approximately 98%.

## 90 Panelize

91 The panelize stage transforms the geocoded data into a panel table with each row representing  
92 a single SME for a given year. Furthermore, it enriches the panel with revenue, expenditure,  
93 and average workforce attributes. Panel tables are larger in size but usually more convenient  
94 for the researchers.

## 95 Discussion and Future Directions

96 The present state of open data in Russia is a matter of serious concern. There is a tendency  
97 for the removal of open datasets, and there is no guarantee that the source datasets required  
98 by this application will remain available in the future. In order to mitigate these concerns, a  
99 decision has been taken to maintain a publicly available backup of the raw source data required  
100 for the tool to function. Nevertheless, there is no explicit commitment regarding the frequency  
101 of these backup operations or the duration of the maintenance period.

102 There are avenues for enhancement of this application, such as the augmentation of geocoding  
103 quality or the incorporation of additional attributes. Any assistance in the development of the  
104 application would be greatly appreciated.

105 For some users, the provision of ready-to-use datasets may be a more appealing option in  
106 comparison to the utilisation of the application. A dataset generated by the tool was published  
107 in November 2024 in the [Research Data Infrastructure portal](#). The future of this portal remains  
108 unclear, therefore, exploring alternative platforms for the publication of this dataset with its  
109 regular updates is being considered.

## 110 Acknowledgements

111 I want to thank Dmitry Skougarevskiy and Ruslan Kuchakov from the [Institute for the Rule of](#)  
112 [Law](#) for their constructive feedback on the potential caveats of working with FTS open data. I  
113 also thank Iuliia Kuzevanova (former coordinator of Research Data Infrastructure, or RDI) for  
114 her assistance in the publication of the dataset generated by this application. Furthermore,  
115 numerous members of the RDI public chat forum have expressed interest in facilitating access  
116 to open data concerning Russian SMEs. This interest was a significant motivating factor in  
117 the completion of the development of this tool.

## 118 References

- 119 Barinova, V., Zemtsov, S., & Tsareva, Y. (2023). *In search of entrepreneurship in Russia.*  
120 *Part i. What prevents small and medium businesses from developing* (p. 400). "Delo"  
121 publishing house of RANEPa. ISBN: 978-5-85006-428-0
- 122 Bondarkov, S., Ledenev, V., & Skougarevskiy, D. (2025). *Russian Financial Statements*  
123 *Database*. <https://doi.org/https://doi.org/10.48550/arXiv.2501.05841>
- 124 Eryomko, I. A., & Goryunova, L. A. (2025). Trends in small and medium-sized business  
125 development in the federal subjects of Russia. *Bulletin of Buryat State University. Economy*  
126 *and Management*, 178(1), 81–88. <https://doi.org/10.18101/2304-4446-2025-1-81-88>

- 127 Gumenyuk, I. S. (2022). On the dynamics of economic activity and its impact on the budgetary  
128 stability of municipalities of the kaliningrad region. *IKBFU's Vestnik. Series: Natural and*  
129 *Medical Sciences*, 1, 44–56.
- 130 Kerimov, A. T., & Mustafaeva, S. R. (2023). Management consulting: Assessment of the  
131 level of territorial concentration of the market of the russian federation. *Scientific Notes*  
132 *of Crimean Engineering and Pedagogical University*, 4(82), 105–109. <https://doi.org/10.34771/UZCEPU.2023.82.4.021>  
133
- 134 Korchagina, I. V. (2025). Potential of small and medium-sized enterprises in containing spatial  
135 inequality of regions of the siberian federal district. *Herald of Omsk University. Series*  
136 *"Economics"*, 23(1), 98–106.
- 137 Kuzora, S. S., & Natarov, I. P. (2022). Digital transformation and big data.  
138 *RUDN Journal of Public Administration*, 9, 150–161. <https://doi.org/10.22363/2312-8313-2022-9-2-150-161>  
139
- 140 Razmanova, S., Volkov, A., & Nesterova, O. (2022). Small and medium business' development  
141 problems (the komi republic case study). *П-Economy*, 15(1), 48–65. <https://doi.org/10.18721/JE.15104>  
142
- 143 Zakharova, K. A., Muravev, D. A., & Zheurova, E. G. (2023). The state and development  
144 of small and medium-sized enterprises in the russian federation. *Tyumen State University*  
145 *Herald. Social, Economic, and Law Research*, 9(3), 233–246. <https://doi.org/10.21684/2411-7897-2023-9-3-233-246>  
146