STI2022
GRANADA | SEP 7+8+9

26th International Conference on Science and Technology Indicators
"From Global Indicators to Local Applications"
#STI2022GRX

Poster

STI 2022 Conference Proceedings

Proceedings of the 26th International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Proceeding Editors

Nicolas Robinson-Garcia
Daniel Torres-Salinas
Wenceslao Arroyo-Machado

Citation: Tsivinskaya, A. (2022).Global Russian Studies: Quantitative Methodologies and the Production of Academic Knowledge. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators*, STI 2022 (sti22189). https://doi.org/10.5281/zenodo.6935745

Collection: https://zenodo.org/communities/sti2022grx/

26th International Conference on Science and Technology Indicators | STI 2022
"From Global Indicators to Local Applications"
7-9 September 2022 | Granada, Spain
#STI22GRX

Global Russian Studies: Quantitative Methodologies and the Production of Academic Knowledge

Angelika Tsivinskaya*

*atsivinskaya@eu.spb.ru
European University at St. Petersburg, 6/1A Gagarinskaya Street, 191187, St. Petersburg (Russia)

Introduction

The aim of our study is the academic expertise of knowledge about Russia through the worldwide search and ranking of academic institutions. Conventionally, one can single out the main journals and conferences on Russian studies. However, studies carried out about Russia are presented in other disciplines (sociology, political science) or fields (higher education studies). Therefore, we decided to analyse not only publications in journals recognised as belonging to the field of Russian studies but also include others that contribute to knowledge about Russia in social sciences and humanities (Castro Torres AF et al. 2022). This gives the unique view of the community of scholars that do not necessarily identify themselves with Russian studies. We can see how they integrate local knowledge into the system of global knowledge production (Hanson et al., 2014). We develop a methodology and implement the collection of the necessary data for ranking institutions in the field of research about Russia from the WoS database. This project will result in a global ranking of academic organizations; methodology for constructing rankings for interdisciplinary areas; scientometric analysis of the field of Russian studies.

We wanted to answer with our study:

- What is written about Russia?
- Who writes papers about Russia?
- Where are located the most successful authors with publications about Russia?

Methodology

To answer those questions, we tried to develop our own methodology consisting of the following steps:

- Seed dataset for keywords
- Selection of keywords
- Forming primary dataset
- Selection of papers
- Cleaning of affiliation information

Seed dataset

Our approach was not just query Initial list of journals was obtained from http://www.slavicreview.illinois.edu/info/related.html. The following list represents a selection of scholarly journals in Russian, Eurasian, and East European Studies. We were able to find ISSN for 44 journals using Scopus journal list as the main source of information. This list represents more it was used to construct possible keywords that could be essential to find research in other areas that potentially can be about Russia. As you can see from the Table 1. The coverage of selected journals was better in WoS comparing to Scopus so we chose WoS as our main source for querying publications. Even though some journals were present only Scopus the greater number of papers were generally indexed in WoS for most of the journals.

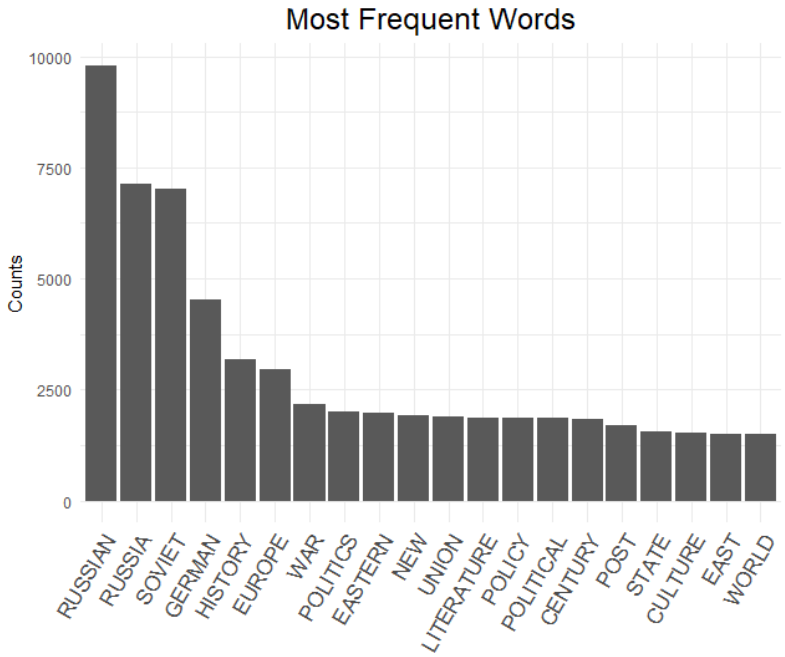
Table 1. Comparison of available records in bibliometric databases.

	WoS	Scopus
ALL	66197	29084
2004-2018	27042	17384

Keywords

We used paper titles for creating keywords (Chen et al., 2016). At first, titles were split into n-grams using simple frequency method. There were 39675 words generated from our seed dataset excluding stop words. We decided to select as potential keywords only nouns and pronouns. Most frequent words in the titles were Russia and Russian as could be expected. Other frequent words are not as suitable as keywords because including them in the query will give too many false positives as they are mostly generic concepts. Therefore, we decided to select words manually from the list of frequent words.

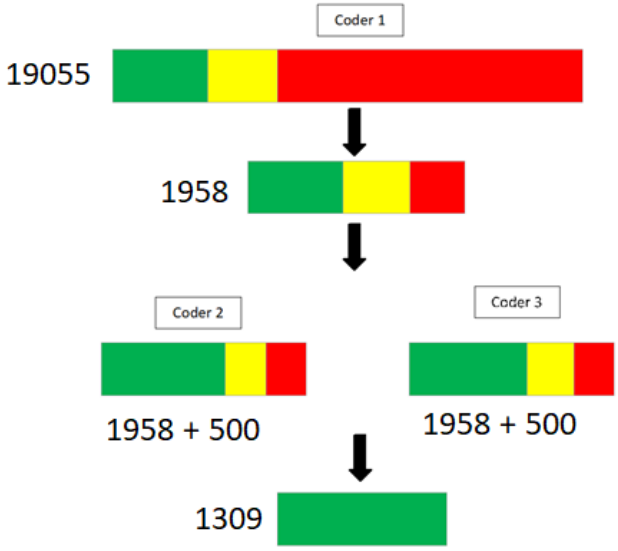
Figure 1: Words extracted from titles of papers.



We choose for more close examination 19055 words which occurred at least 2 times in titles. Coder 1 looked through the entire list of these words and selected 1958 words as possible keywords. Subsequently, Coder 2 and Coder 3 looked through words that were selected by Coder 1 and in addition each one has 500 words that were randomly selected from words that Coder 1 excluded to test for consistency of exclusion process. This scheme was employed to reduce costs spend for checking keywords and at the same time guarantee the possibility to see

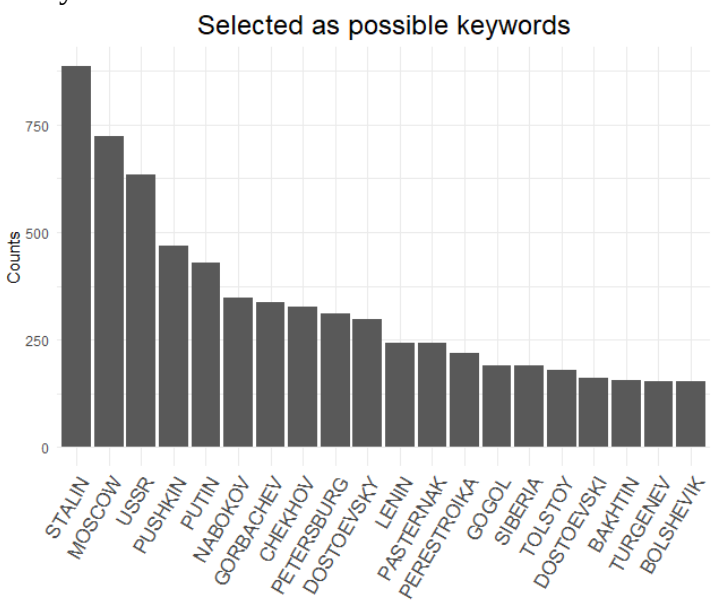
variability in selected keywords. We also calculated for each possible keyword a percentage of titles where it co-occured with keywords RUSSIA*. This indicator was useful in several instances such as to flag words that definitely unrelated to our area of interest as well as to see words that have strong association with Russia that in not needed to be referred in the title directly.

Figure 2: Scheme for manual selection of keywords.



Here we closely examine the consistency of selected keywords between coders. As you can see from the Table 3. coders have quite different perception of uncertainty about selection. Coder 1 is sure in fewer words to be selected as keywords. Coder 2 has the highest number of words that were selected as keywords. There is a higher degree of agreement concerning randomly assigned excluded words. Overall, 1309 words were selected as keywords. For all preselected words, total agreement was only in 34.4% but for two or all coders was 97.7%. We also wanted to see if the agreement on selected keywords depends on their frequency (whether for words that are more frequent is easier to reach the consensus).

Figure 3: Selected as keywords.



Primary dataset

We restricted results of our query by English languages and article as document type. At first, we consider the period from 2004 to 2019. The number of papers in the last five years is roughly the same as the ten previous so we decided to focus on the later period from 2014 to 2019.

Table 2. Number of publications in different subsets.

Subset	N	In both
Publications selected with keyword RUSSIA* in TOPIC	28 984	-
Publications selected with keyword RUSSIA* in TOPIC and Publications selected with other keywords in TITLE	41 370	4 998
Publications selected with keyword RUSSIA* in TOPIC and publications selected with other keywords in TOPIC	78 753	11 631

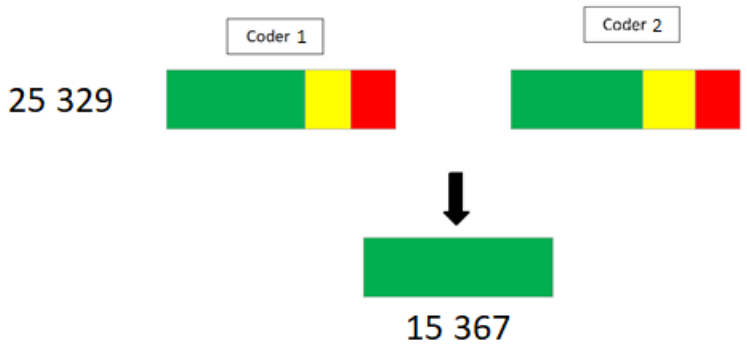
Coding abstracts

The selection by keywords is quite coarse approximation for our area and it could have a high number of false positives that is why we needed to manually check papers that should be in our final dataset. Additionally, as we are focused only on social sciences and humanities, we excluded paper that were classified outside our domain.

To estimate time coast for manually selected papers for our further analysis and facilitate coding process for the next step we randomly sampled of 2500 papers for manual coding. We also wanted to see which subset of papers is more suitable for our research. In the random sample, 513 papers were coded as belonging to studies about Russia and 815 were not. For rest of papers the decision about selection was inconclusive just by reading only titles and need additional examination. 26.3% about Russia which are not selected with just keyword RUSSIA*. 15.8% is not about Russia which are selected with keyword RUSSIA*. Most of them have RUSSIA in the abstract. For the sake of time, we can assume that papers with RUSSIA in title are about Russia.

The whole set N=25 329 was viewed by two coders including reading not only titles but as well abstract for difficult cases. We were interested in comparing the degree of agreement between coders in two parts of our database. One contained keyword Russia* in abstract or keywords both author and WoS generated (N=7771) and the second part is based on our selected keywords (N=17558). Coder 1 selected 4997 publications as belonging to our field of interest whereas Coder 2 selected 8 126.

Figure 4: Scheme for manual selection of publication.



Our primary dataset included all papers that had Russia* keyword in the title and papers that were selected at least by one coder. The total sample size N=15367 publications.

Cleaning affiliations

There was no affiliation information for N=1130 authors in the selected papers. In most cases, the affiliations were missing for all authors of papers and in rare cases; information was not available just for subset of authors of one paper. We combined results of manual check for affiliation information with suggested disambiguation by package refplirt in R.

Organizational name unification

The dataset from WoS contained affiliation records presented as they were written by authors. We go through additional step of unification of possible variant for names of organization as some authors have alternative ways of writing their affiliation information, especially in case of organization from Russia where several possible translations can be used even if organization has a guidance. For this task, first we list all organization names in our dataset and then decided to merge variants if this particular name is used at least by 5 authors of papers. Other rarer organization names were omitted for the sake of coding time. We ended up N=401 organizational names from Russia and N=633 form other countries.

References

Chen, Guo; Xiao, Lu (2016). Selecting publication keywords for domain analysis in bibliometrics. A comparison of three methods. In *Journal of Informetrics* 10 (1), pp. 212–223.

Hanson, Stephen E.; Ruble, Blair A. (2014). Rebuilding Russian Studies. In *Problems of Post-Communism* 52 (3), pp. 49–57.

Castro Torres AF, Alburez-Gutierrez D. (2022). North and South: Naming practices and the hidden dimension of global disparities in knowledge production. In *Proc Natl Acad Sci USA*. 9 (10)