# Russian and International Data Sources: Integration of Data on Russian Research Organizations

Zinaida V. Apanovich

*A.P. Ershov Institute of Informatics Systems, Siberian Branch, Russian Academy of Sciences, Lavrentieva pr., 6, Novosibirsk, 630000*

### Abstract

This paper considers international and Russian-language data sources providing information about Russian research-related organizations. Information about research organizations is an important attribute that enables one to identify the authors of scientific publications, as well as to analyze the geographical distribution of publications and to assess the impact on the citation of the publications associated with geographic factors. However, information about national research organizations, for example, information about Russian research organizations, is often incomplete or distorted in international databases. Data sources such as GRID, Russian and English chapters of Wikipedia, Wikidata and eLIBRARY.ru are considered. It is demonstrated that Russian-language data sources contain more information about Russian research-related organizations than most international data sources, but this information is not available in English-language data sources. To solve this problem, a method for integrating information from multilingual data sources has been developed. Experiments on the comparison and integration of information about Russian research organizations in international and Russian data sources are outlined. An experimental version of the database of scientific organizations comprising 3143 scientific organizations has been created. The work is an intermediate step towards the creation of an open and extensible knowledge graph.

### Keywords

Knowledge graph, multi-lingual knowledge graphs, identity resolution, research-related organizations, correctness

## 1. Introduction

Information on research organizations is an important attribute that enables the identification of the authors of scientific publications, as well as the analysis of the geographical distribution of publications and assessment of the impact on the citation of the publications associated with a geographic factor [3]. Regrettably, for example, information about Russian research organizations, is often incomplete or distorted in international databases.

One of the largest international open databases of scientific organizations is GRID [4] (Global Research Identifier Database, https://www.grid.ac/). GRID is a free and openly accessible global database of research-related organizations, cataloging research-related organizations and providing each of them with a unique and persistent identifier. Its data is downloadable as Excel, JSON or RDF (ttl format) files. This database contains information on more than 102,390 research-related organizations from 220 countries. The GRID data are integrated into the SN SciGraph knowledge graph developed by Springer (https://www.springernature.com/gp/researchers/scigraph).

The information on the organizations presented in GRID includes their postal address (100%), geographic coordinates (longitude and latitude) (99%), and the URL (90%). Each organization is provided

SSI-2021: Scientific Services & Internet, September 20–23, 2021, Moscow (online)
EMAIL: apanovich_09@mail.ru (Z.V. Apanovich)
ORCID: 0000-0002-5767-284X (Z.V. Apanovich)

with such attributes as *geonames_city_id*, *geonames_country_id*, *geonames_country_code*, etc. Thanks to the use of GRID data, all information about publications stored in the SN SciGraph is geo-referenced.

Also, GRID maintains links to global bibliographic resources such as ROR (Research Organization Registry, https: //ror.org), Crossref https://www.crossref.org/, and ISNI (International Standard Name Identifier, https: //isni.oclc.org/). GRID currently contains data on 2066 Russian research organizations.

However, the information pertaining to Russian research organizations is incomplete and contains obvious inaccuracies. For example, GRID has a page dedicated to the Siberian Branch of the Russian Academy of Sciences (SB RAS, https://www.grid.ac/institutes/grid.415877.8). This page indicates that “Институт космофизических исследований и аэрономии им. Ю.Г. Шафера Сибирского отделения Российской академии наук” is the Russian appellation for “SB RAS.” In actual fact, this is the Russian equivalent of the “Shafer Institute of Cosmophysical Research and Aeronomy” (https://www.grid.ac/institutes/grid.435157.1).

Also, along with several institutes formally related to the SB RAS, some educational organizations of different subordination are listed as the subsidiary organizations (“child institutes”) of the SB RAS. For example, the East-Siberian Institute of the Ministry of Internal Affairs of the Russian Federation, (https://www.grid.ac/institutes/grid.445063.0), Siberian Law Institute of Russian Federal Drug Control Service (https://www.grid.ac/institutes/grid.445537.4), etc. are mentioned by GRID as “child institutes” of the SB RAS.

Since GRID provides links to global bibliographic resources such as ROR, Crossref, and ISNI, it is interesting to find out whether the information about Russian organizations presented on these international platforms differs from that in GRID. Regrettably, our experiments have shown that ROR copies information about Russian research-related organizations, either correct or erroneous, from GRID. For example, the outdated information on the website of the no longer existing Novosibirsk Humanitarian Institute presented in GRID (https://grid.ac/institutes/grid.445355.6) is also found in ROR (https://ror.org/00nnwpb90).

A more important example of a copied error is the link between the GRID page dedicated to the Siberian Branch of the Russian Academy of Sciences (https://www.grid.ac/institutes/grid.415877.8) and the ID of this organization in ROR, https://ror.org/02frkq021, which gives two “equivalent” names of this organization: “SB RAS” and “Institute of Space Research and Aeronomy Named after Yu.G. Shafer of the Siberian Branch of the Russian Academy of Sciences”.

Both GRID and ROR present incomplete data; for example, none of them contains information about the A.P. Ershov Institute of Informatics Systems SB RAS, nor of many other Russian scientific organizations.

There is a striking discrepancy in the number of Russian organizations in GRID and the relevant number in the largest database of Russian research-related organizations eLIBRARY.ru [5].

These examples suggest that Russian-language data sources contain more complete and correct information on Russian organizations than their English-language counterparts. The largest Russian-language data sources on Russian organizations are eLIBRARY.ru and Russian Wikipedia. It is reasonable to compare the organizations shown in GRID and in Russian Wikipedia.

These examples signal the need for the integration of the information contained in international and in Russian data resources. To solve this problem, a method for integrating information from multilingual data sources has been developed.

## 2. GRID and Russian data sources: data comparison

### 2.1. Grid and Wikipedia

The GRID database maintains links to the pages of Russian organizations in the English-language Wikipedia. Although these links look natural to an English speaking user, it would be even more natural to search for the information on Russian organizations in the Russian-language Wikipedia. At the time of our experiments, GRID contained 2019 pages of Russian organizations, only 412 of which had links to the pages in the English-language Wikipedia. Among these 412 pages, 398 pages were related by interlanguage links to the pages of the Russian-language Wikipedia. Predictably, the Russian-language Wikipedia contains more information about the Russian organizations presented in GRID. For example, the GRID page devoted to the Federal Agency for Scientific Organizations (FASO, https://www.grid.ac/institutes/grid.484124.f) states that FASO was established in 2013, the link "Institute Links" (https://fano.gov.ru/en/) claims that this address cannot be reached, and a link to the English Wikipedia is not available. However, there is a relevant page in the Russian Wikipedia (https://ru.wikipedia.org/wiki/Федеральное_агентство_научных_организаций) stating that this organization was abolished on May 15, 2018. The same information is duplicated in the Wikidata dataset (Federal Agency for Scientific Organizations, Q16711297) but GRID does not show this page.

To test our hypothesis, we extracted a list of Russian research organisations from GRID, together with such attributes as the English name of an organization, its Russian name, acronyms, aliases, link to the organization's web-site, link to the organization's page in the English Wikipedia, city and country. The pages of only 412 organizations of all the Russian research organizations presented in GRID had a link to a page in the English version of Wikipedia, and 398 of them had a link to a Russian Language page in Wikipedia.

Then, we used the data extracted from GRID to search for the appropriate organizations in the Russian version of Wikipedia by means of Wikipedia_API.

The attributes used included the URL of the organization's website, its English-language name and Russian-language names, etc. Even if not highly efficient, this search produced another 674 pages found in the Russian-language Wikipedia. Among them, 353 Russian pages were linked by cross-language links to the English-language Wikipedia.

In total, 835 matchings between the Russian Wikipedia and GRID pages were found. Thus, this experiment has shown that though the Russian-language version of Wikipedia stores much more information about Russian scientific organizations than the English-language version, this information remains inaccessible to the English-language databases.

The main problem was that explicit links are not many, and the search for the names of organizations is complicated because different databases contain different names of the same organizations. We plan to improve the existing imperfect matching algorithm.

## 2.2. GRID and eLIBRARY.ru

eLIBRARY.ru (Q4037789) it is the leading electronic library of scientific periodicals in Russia in the world. It stores data on science, technology, medicine and education and includes information on over 34 million publications, more than 1 million researchers, and over 12, 000 organizations. Contrast with the number of Russian organizations stored in GRID (2066) is striking. What is the reason for the big difference? It is easy to see that only a part of the eLIBRARY.ru -listed organizations are research-related. This list contains all federal ministries and bodies subordinated to these ministries, regional administrations, commercial organizations, banks, hospitals, individual entrepreneurs, etc. For example, it is possible to find House-Building Plant No. 7, a company having neither publications nor references; the only information on this organization is its postal and legal address. In total, about one third of the list of the organizations stored by eLIBRARY.ru (4505 organizations) is not related to publication activity. Just like GRID, eLIBRARY.ru contains many descriptions of no longer existing organizations.

Each organization in eLIBRARY.ru is described using such attributes as the full name of the organization in Russian and in English, the Russian and English acronym, country, region, Russian and English name of the city in which the organization is located, postal address in Russian, Russian and English postal address, legal address, parent organization, type of organization, fax, email and web-site URL. Each organization has a unique identifier. For example, the identifier of the Ershov Institute of Informatics Systems is 593, and that of the SB RAS is 2378. There are no links to global bibliographic data sources in eLIBRARY.ru. In order to compare the data on the organizations listed in eLIBRARY.ru. and GRID, a program was written. It discovered only 709 matchings between the eLIBRARY.ru and GRID pages describing Russian research-related organizations. The main reason for the small number of matchings is the spelling difference in the names of organizations given in two different data sources.

Currently, there is a data source that tries to integrate information about organizations from all language chapters of Wikipedia. Moreover, it collects all the identifiers assigned to the organizations by global organizations. This data source is Wikidata (wikidata.org).

## 3. eLIBRARY.ru and Wikidata

An example of a highly promising international data source is Wikidata.org. Wikidata emerged in 2014 [6] as a structured data source for fact management in various language versions of Wikipedia. The Wikidata's developers plan to make it the central management platform for Wikipedia, integrating data from all Wikipedia language "chapters".

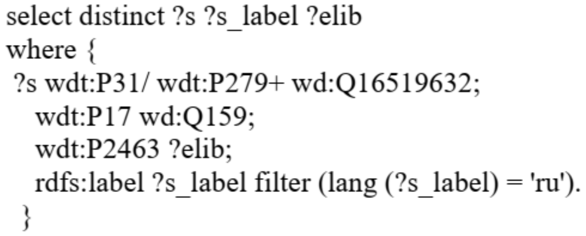
To integrate data, each entity is assigned an identifier independent of a specific language version, and all statements concerning this entity and found in all language versions of Wikipedia are combined. Like GRID, Wikidata supports links to global bibliographic resources by specifying the identifiers of organizations in these data sources. In particular, the following data sources are indicated in wikidata.org: Virtual International Authority File database (VIAF ID, property P214), Library of Congress authority ID (authority ID, property P244), GRID.ac global research identifier database ID (property P2427), ROR Research Organization Registry ID (property P6782), Russian organization number (property P7011), ISNI International Standard Name Identifier ID (property P213), eLIBRARY.ru organization ID,  (property P2463), and Crossref funder ID (P3153).

Wikidata also contains the short names of an organization in its native language and in English, information about its type and geographical location (country, region), dates of inception and closure. For example, the SB RAS page (https://www.wikidata.org/wiki/Q3032414) provides alternative names of this institution in 14 languages. However, for unknown reasons, its official name (P1448) is shown in the Belorussian language and the list of its subsidiaries (P355) contains the same mistakes as the list of child institutions shown on the corresponding GRID page.

Another example is the A.P. Ershov Institute of Informatics Systems (Q4201722), which is not considered as a subsidiary of the SB RAS. Besides, the Wikidata page indicates that the institute was named after Alexandra Petrovna Ershova (Q60830445), a Russian theater teacher rather than Academician Andrey Ershov (Q1961494), Russian computer scientist. Besides, the image one can see at this page (provided by wikidata.org) has evidently nothing to do with Academician Andrey Ershov, while the photo provided by the corresponding Russian – language page in Wikipedia is true.

Numerous facts of this kind point to the need to establish correspondence between data, compare and verify the difference between Russian and international data sources.

Despite the existence of a special property describing the identifier of an institution in eLIBRARY.ru, few Russian institutions represented in Wikidata.org have eLIBRARY.ru identifiers. So, by running a SPARQL query looking for scientific organizations (wd: Q16519632) located in Russia and having a eLIBRARY.ru identifier specified in wikidata.org, we obtained a list of eighty-six institutions, mainly educational. Novosibirsk State University, for example, has the eLIBRARY.ru identifier 214. An example of a query retrieving scientific organizations from the Wikidata website with an identifier of eLIBRARY.ru is shown in Figure 1.



```
select distinct ?s ?s_label ?elib
where {
 ?s wdt:P31/ wdt:P279+ wd:Q16519632;
   wdt:P17 wd:Q159;
   wdt:P2463 ?elib;
   rdfs:label ?s_label filter (lang (?s_label) = 'ru').
 }
```

**Figure 1:** SPARQL query retrieving scientific organizations with the identifier of eLIBRARY.ru in wikidata.org

Table 1 shows the number of entities Organizations and Scientific Organizations that have identifiers in various global bibliographic data sources. For example, only four institutions out of 274 subordinates to the Russian Academy of Sciences (wd: Q4201890) have a eLIBRARY.ru identifier, which means that the task of comparing data in the above-mentioned sources is very relevant. Below we are going to dwell on our approach to solving this problem.

**Table 1.**
The number of entities *Organizations* and *Scientific Organizations* having identifiers in various global bibliographic data sources

| Data sources and the corresponding Wikidata properties | Number of Scientific organizations (Q1651963, 21473 in total) having a wikidata property | Number of Organizations (Q43229, 34911 in total) having a wikidata property |
|---|---|---|
| GRID (P2427) | 989 | 1535 |
| eLIBRARY.ru (P2463) | 86 | 103 |
| OGRN (P7011) | 863 | 1186 |
| VIAF (P214) | 769 | 2094 |
| Library of Congress (P244) | 648 | 1538 |
| ROR (P6782) | 982 | 1527 |
| ISNI (P213) | 714 | 717 |

Note that as the data stored in Wikidata is incomplete, it is currently impossible to obtain all reliable information using the SPARQL query. For example, a SPARQL query searching for all Russian scientific organizations returns only the organizations explicitly stating that they are located in Russia. If the requirement of Russian affiliation is made optional (OPTIONAL), the resulting data set includes some organizations without any information about the country of their origin. Therefore, the problem of matching and comparing entities described in different data sources needs to be solved programmatically.

The input of the data integration program is the table *Organizations*. Each row of this table corresponds to an institution listed in eLIBRARY.ru. The columns of the table *Organizations* correspond to such attributes of eLIBRARY.ru as the full name of an organization in Russian, its name in English, Russian abbreviation of the name, English abbreviation of the name, country, region, Russian name of the city were the organization is located, English name of the city, postal address of the organization in Russian and in English, its legal address, its parent organization, type of the organization, its fax, and official web site.

The integration algorithm results in an extended table showing whether the description of an organization was found in Wikidata. In case of a positive result, the corresponding row of the table is supplemented with information extracted from Wikidata. In particular, the Wikidata_name of the organisation, Wikidata_identifier, Wikidata_alias, Wikidata_year of foundation, and international identifiers, such as VIAF_Id, eLIBRARY_Id, GRID_Id, are added. The options to the negative result are “organization not found” and “there is not enough data to identify the organization”.

The integration algorithm is structured as follows:

1. Pre-processing the names of organizations (translation into lower case, deletion of words such as "ЗАО" (CJSC, closed joint-stock company), "ООО"(private limited liability company), ОАО (OJSC, open joint-stock company), “им”, “имени" (named after), etc., cyclic replacement of some words using a dictionary of synonyms). For example, the words “RF”, “Russia” and “the Russian Federation” in the name of an institution should be considered as synonyms. Also, the words “mayor's office”, “administration” and “government” are often used interchangeably in the names of organizations.
2. API Wikidata-based search for entities by one of the names of an organization specified in the table (4 variants of the name, transformed names, URL). If the search is successful, a JSON file is returned with brief information about the element found. This data allow for extracting additional information about the entity: its name, identifier, entity type, all the names available. Wikidata contains information about all available names of an entity in the “Also known as” field.
3. Checking whether an entity found in Wikidata is equivalent to an institution from eLIBRARY.ru. To do this, the coefficient of matching between all the variations of full and transformed names of the entity is calculated, the URLs of the organizations are compared, the type of the entity is found, the country where it is based and location inside the country are checked.
4. Supplementing the original table with information from Wikidata.

Step 3 of the integration algorithm is the most difficult and consists of the following stages.

**3.1 Comparing the names of organizations**. When two lists of names of an organization extracted from eLIBRARY.ru and Wikidata are compared, a string similarity of these names is estimated. First, each pair of strings is tested for complete textual coincidence. In case of a negative result, each string is divided into separate words, which are transferred to the nominative form using the module of morphological analysis. For the resulting rows, the matching coefficient is calculated using the following formula:

$$matching\ coefficient = 2 \cdot \frac{number\ of\ matching\ words}{total\ number\ of\ words}$$

The best matching coefficient is calculated for all the name variants. If for a pair of names this coefficient is greater than 0.68, the result is considered positive, and the rest of the attributes are compared.

**3.2 Comparing the URLs of organizations**. When comparing the URLs, we noticed that their references may differ from a database to database. For example, the URL of Administration of the Arkhangelsk Region in eLIBRARY.RU is http://www.dvinaland.ru/, and in Wikidata it is https://dvinaland.ru/. Thus, the difference between the “http” and “https” constructions and the presence of the “www” construct in one of the addresses will make the result of the comparison negative. To avoid such situations, the constructs specified are removed when the URLs are compared.

**3.3 Checking the type of the object found**. You need to make sure that the object found is indeed an organization. Using a SPARQL query, we generated a CSV file containing the names of all subclasses of the Organization class in the Wikidata ontology. The type of the entity found is compared with the elements of this file. If the comparison is negative, the entity is rejected.

**3.4 Checking the location of the organization.** If Russia is not specified as the country where the institution is based, the search stops and the database records that the organization is not located in Russia. If Russia or another country is not indicated on the Wikidata page, information about the location of the headquarters is extracted. As a rule, Wikidata indicates the city and administrative-territorial unit. Note that Wikidata may indicate the Moscow region instead of Moscow, for example, which can lead to an incorrect comparison result. To solve this problem, we used information about a hierarchy of geographic objects. We created a JSON file that includes the distribution of all Russian cities by regions. The program extracts the name of the city from Wikidata; if it does not match the location specified in eLIBRARY.RU, the name of the city will be replaced with the name of the region in which the city is based. Subsequently, the geographic locations are compared again. If the locations do not match, the organization is considered incorrect and the program goes to the next institution.

Thus, at the moment, an organization is considered to be correctly identified in two cases:

1. Site URLs and some pairs of name variations match.
2. The entity is an organization, the names of the organization in the two sources coincide, information about the sites is not complete, the organization is located in Russia.

For all organizations recognized by the algorithm as identical, information is combined based on an extension of the schema.org ontology. Currently, correspondence has been established between 3143 organizations in Wikidata and eLibrary.ru. The resulting experimental data source will be further expanded and integrated with other data sources, such as GRID.

## 4. Conclusion

Experiments with English-language and Russian-language data sources have shown that Russian-language information sources contain more information about Russian-speaking scientific organizations. Regrettably, this information remains largely inaccessible to English-language data sources. To solve this problem, a method for integrating information from multilingual data sources has been developed. An experimental version of the database of scientific organizations comprising 3143 scientific organizations has been created. It is planned to turn this base into an open and extensible knowledge graph. The authors also believe that in order to maintain the completeness and correctness of information about scientific entities, each scientific organization should maintain its own page on international platforms, which would indicate all the identifiers of the organization.

## 5. Acknowledgements

The author thanks Yurieva I.O. and Chvyrova O.S. for participation in the implementation of various versions of the integration program.

## 6. References

[1] Z. Apanovich, Matching of authors and publications in multilingual bibliographic knowledge bases, in: CEUR Workshop Proceedings. SSI 2019, Proceedings of the 21st Conference on Scientific Services and Internet, 2020, pp. 26–37.

[2] A. Haira, V. Radevski, K. Tochtermann, Author profile Enrichment for Cross-linking Digital Libraries. Research and Advanced Technology for Digital Libraries Springer International Publishing. Lecture Notes in Computer Science 9316 (2015) 124–136. https://doi.org/10.1007/978-3-319-24592-8_10.

[3] A. Manocci, F. Osborne, E. Motta, Geographical trends in academic conferences: An analysis of authors' affiliations. Data Science 2 (1) (2019) 181–203. https://doi.org/10.3233/DS-190015.

[4] Global Research Identifier Database. URL: https://www.grid.ac/.

[5] Scientific online library eLIBRARY.ru. URL: https://www.elibrary.ru/

[6] A. Ismailov, D. Kontokostas, S. Auer, J. Lehmann, S. Hellmann, Wikidata through the Eyes of DBpedia. URL: http://www.semantic-web-journal.net/system/files/swj1462.pdf