ОБЩИЕ ВОПРОСЫ МЕТОДОЛОГИИ СЕТЕВОГО АНАЛИЗА

DOI: 10.19181/4m.2023.32.1.1
EDN: ZBAAGN

Д.В. Мальцева, И.А. Павлова,
Л.В. Капустина, В.А. Ващенко
(*Москва*)
Д. Фиала
(*Чехия*)

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ВОЗМОЖНОСТЕЙ WOS И ELIBRARY ДЛЯ АНАЛИЗА БИБЛИОГРАФИЧЕСКИХ СЕТЕЙ[1]

Дарья Васильевна Мальцева – кандидат социологических наук, заведующая Международной лабораторией прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. Email: dmaltseva@hse.ru.
Ирина Анатольевна Павлова – кандидат экономических наук, старший научный сотрудник Международной лаборатории прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. Email: iapavlova@hse.ru.
Лика Владимировна Капустина – стажер-исследователь Международной лаборатории прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. Email: lkapustina@hse.ru
Василиса Андреевна Ващенко – стажер-исследователь Международной лаборатории прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. Email: vvashchenko@hse.ru.
Далибор Фиала – доцент Факультета прикладных наук Департамента компьютерных наук и инженерии, Западночешский университет, Чехия, Пльзень. Email: dalfia@kiv.zcu.cz.

[1] Статья подготовлена в ходе проведения исследования в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ).

В статье проводится сравнительный анализ баз данных научных публикаций Web of Science Core Collection и eLibrary с целью выделения их особенностей и описания возможностей анализа при изучении библиографических сетей российских авторов. Актуальность исследования определяется необходимостью адаптации и разработки подходов и инструментов для сбора, предобработки и анализа библиографических данных на русском языке. Анализ проводится на основе сравнения массивов данных публикаций в научных журналах в области социологии, выгруженных за период 2010–2021 гг. Выделяются основания для сопоставления двух баз, характеризующие получение доступа к данным, организацию данных в базах, количественные и содержательные характеристики данных. Анализ отобранных параметров позволяет найти пересечения между массивами данных и содержательными результатами анализа. Делаются выводы о соотношении двух баз, их возможностях и ограничениях по использованию в качестве основного (единственного) источника информации, даются рекомендации об их использовании для изучения отечественной науки.

Ключевые слова: сетевой анализ, сравнительный анализ, библиографические базы, библиографические сети, eLibrary, Web of Science

Введение

Анализ библиографических сетей – частный случай применения методологии анализа социальных сетей. Он основан на построении и анализе сетей соавторства и коллаборации, цитирования и социтирования, библиографического сочленения, соприсутствия библиометрических единиц анализа. Направление способно показать закономерности развития взаимодействия в научном сообществе, определить его структуру, динамику, направления исследований [1; 2; 3]. Основные этапы исследования с применением анализа библиографических сетей подразумевают использование технологических решений для 1) формирования базы библиографических данных, 2) ее предобработки и постро-

ения различных видов библиографических сетей и 3) последующего изучения с применением методов сетевого анализа (social network analysis).

Как и в любом исследовании, выбор источника информации является определяющим для качества анализа – по принципу GIGO[1], получение достоверных результатов напрямую зависит от стратегии поиска и полноты используемой базы данных. Выбор баз данных для исследователя является достаточно широким – помимо часто используемых для учета эффективности работы ученых баз научного цитирования Web of Science (WoS) и Scopus, большую популярность приобрели бесплатные базы данных, агрегирующие библиографическую информацию, такие как «универсальные» Google Scholar и OpenAlex, включающие информацию о патентах Digital Science Dimensions и Lens, базы медицинских исследований PubMed и Cochrane, научные социальные медиа SciFinder, Mendeley и др. Значительное количество баз данных научных публикаций привело к появлению исследований, посвященных сравнительному анализу различных площадок, где они сравниваются по различным характеристикам.

Международные базы данных могут выступать источниками информации и при изучении научного производства российских авторов. Данные из WoS и Scopus до недавнего времени[2] активно использовались для оценки научной продуктивности ученых в рамках ряда государственных программ финансирования университетов (например, «Проект 5–100»). Однако механика формирования этих баз предполагает неполное покрытие всей

[1] GIGO – аббревиатура для используемой в информатике фразы “garbage in, garbage out”, означающей, что при неверных входящих данных будут получены неверные результаты, даже если сам по себе алгоритм правилен. В русском языке аналогом является пословица «что посеешь, то и пожнешь».

[2] До момента прекращения работы данных организаций в России в связи с началом проведения СВО в 2022 г.

научной продукции отдельной страны: представленность индексируемых в базах журналов (и издательств) является выборочной и основана на наукометрических показателях. В связи с этим при использовании международных баз данных в качестве источника объем производимой в России научной литературы оказывается недооцененным.

В поле российской науки имеется несколько баз данных, аккумулирующих информацию о научных публикациях отечественных исследователей. Крупнейшей в России базой научных публикаций является научная электронная библиотека eLibrary, которая интегрирована с Российским индексом научного цитирования (РИНЦ) – созданной по заказу Минобрнауки РФ общедоступной базой данных российских научных публикаций. В качестве альтернативы можно отметить научную библиотеку КиберЛенинка, предоставляющую доступ к публикациям на основе принципов открытой науки. Эти базы данных аккумулируют значительное число публикаций, производимых российскими авторами (в том числе и через интеграцию с международными базами научного цитирования и получение информации о публикациях в международных журналах[1]), и могут рассматриваться как источники данных для изучения российской науки.

В поле библиометрического анализа разработаны специальные инструменты, которые позволяют производить анализ библиографических сетей на основе данных из международных наукометрических баз: производить предобработку полученных данных, строить различные виды сетей и затем анализировать их с помощью методологии сетевого анализа. Пакеты Bibliometrix для R и Python и их веб-приложение Biblioshiny[2] позволяют

[1] Авторы не имеют данных о том, продолжается ли эта интеграция с 2022 г. по настоящее время.

[2] Bibliometrix – An R-tool for comprehensive science mapping analysis [site]. URL: https://www.bibliometrix.org/home/ (date of access: 08.04.2024).

работать с базами данных Scopus, WoS, PubMed, Digital Science Dimensions, Cochrane, Lens и OpenAlex для анализа социтирования, библиографического сочленения библиографических единиц анализа, соавторства и соприсутствия ключевых слов. Программа VOSviewer[1], помимо перечисленных, позволяет работать с такими базами, как Crossref, Europe PMC, Semantic Scholar, OpenCitations и WikiData через их API-сервисы, запрашиваемые в интерактивном режиме в самой программе (осуществляя таким образом и сбор данных). Программа CitNetExplorer[2], предназначенная для анализа цитирований научной литературы, импортирует данные из WoS. На использование данных WoS ориентирован и методологический подход, разработанный В. Батагелем, А. Ферлигой и П. Дореаном (подробнее о подходе см.: [4]), применявшийся для анализа некоторых зарубежных научных дисциплин и описанный в отечественной литературе [5, 6], который использует программу WoS2Pajek для создания из данных WoS сетевых файлов для работы в программе для анализа и визуализации больших сетей Pajek[3].

В связи с тем, что указанные инструменты используют англоязычные коллекции словарей для предобработки и нормализации данных (дизамбигуации имен, лемматизации и токенизации слов), они могут применяться для анализа только части работ российских авторов, опубликованных в международных базах на английском языке. Если же речь идет о публикациях на русском языке, то использование этих инструментов затруднено и должно сопровождаться их адаптацией. Единственным инструментом, позволяющим измерять публикационную активность ученых и ор-

[1] VOSviewer. Visualizing scientific landscapes [site]. URL: https://www.vosviewer.com/ (date of access: 08.04.2024).

[2] CitNetExplorer. Analyzing citation patterns in scientific literature [site]. URL: https://www.citnetexplorer.nl/ (date of access: 08.04.2024).

[3] Pajek. Analysis and visualization of very large networks [site]. URL: http://mrvar.fdv.uni-lj.si/pajek/ (date of access: 08.04.2024).

ганизаций в русскоязычном научном пространстве, является информационно-аналитическая надстройка Science Index, реализованная на платформе eLibrary и основанная на анализе внесенных в базу публикаций. Однако ни функции сбора данных, ни инструменты для сетевого библиометрического анализа на площадке не представлены. Таким образом, в отечественном научном пространстве отсутствует полноценная методология по сбору, предобработке и анализу библиографических данных на русском языке.

В текущей ситуации, связанной со сложностями применения привычных инструментов оценки публикационной продуктивности российских авторов, актуализируется задача по изучению различных баз данных с точки зрения их возможностей для наукометрического анализа. В данной статье мы фокусируемся на сравнительном анализе двух научных баз – международной базы WoS и российской базы eLibrary – для изучения возможностей, которые они предоставляют для анализа библиографических сетей российских авторов. Тогда как использование базы WoS как источника данных позволяет использовать ряд предложенных методологических решений по обработке и анализу данных, выбор базы eLibrary в качестве источника данных предполагает необходимость адаптации существующих и разработки новых технологических решений. Сравнение двух баз проводится посредством сравнения их контента – формата, полноты представления и объема библиографических данных по публикациям российских авторов (насколько похожи данные), а также содержательных характеристик, получаемых при сетевом библиографическом анализе (насколько похожи результаты анализа). В качестве примера взяты все публикации в области социологии за 2010–2021 гг. на обеих площадках – массивы данных из 3995 публикаций в WoS и 75 232 публикаций в eLibrary.

В результате литературного обзора выделяются параметры для сравнения двух баз данных, которые затем анализируются посредством описательного, статистического и сетевого анализа.

Статистический и сетевой анализ основных библиометрических единиц (публикаций, авторов, соавторов, журналов, ключевых слов) и базовых производных сетей для обоих массивов позволяет сравнивать распределения и рейтинги изучаемых библиографических единиц и делать выводы о содержательных различиях между массивами данных. Анализ позволяет делать выводы о соотношении двух баз данных, их возможностях и ограничениях по использованию в качестве основного (единственного) источника информации и давать рекомендации об их использовании для изучения отечественной науки.

Сравнение баз данных научных публикаций

Платформа WoS компании Clarivate Analytics является первой базой научного цитирования, построенной на основе Индекса цитирования научных статей (Science Citation Index), разработанного в 1960-е гг. одним из основателей наукометрии Ю. Гарфилдом. На основе анализа цитирований статей Гарфилд разработал подход к рейтингованию научных журналов, составляющих «ядро» научных дисциплин (Core Collection – CC), в котором позже появились и другие коллекции научных публикаций[1]. Долгое время WoS имела монополию на предоставление информации о научной литературе, однако с появлением в 2004 г. платформ Scopus и Google Scholar ситуация изменилась. Scopus, как и WoS, стал предоставлять информацию о цитировании, получаемую в виде метаданных от производителей научной литературы, однако существенно расширил покрытие научных журналов. Google Scholar расширил диапазон источников до материалов конфе-

[1] Индексы цитирования социальных наук (Social Sciences Citation Index – SSCI), искусств и гуманитарных наук (Arts and Humanities Citation Index – AHCI), новых источников (Emerging Sources Citation Index – ESCI), конференционных публикаций и книг.

ренций, книг, диссертаций, отчетов и других типов публикаций с сайтов издателей и конференций, используя автоматические методы извлечения информации из электронных файлов научных публикаций. С 2004 г. появилось много других агрегаторов научной информации, таких как OpenAlex (универсальная база), Digital Science Dimensions и Lens (начинались как патентные базы), PubMed и Cochrane (медицинские исследования), SciFinder, Mendeley, ResearchGate (научные социальные медиа) и др.

В обзорах развития наукометрических исследований [1; 2] приводится масса ссылок на исследования, сравнивающие базы данных WoS, Scopus и Google Scholar друг с другом, а также с более новыми базами. Рассмотреть все эти публикации не представляется возможным ввиду их большого количества, однако отметим ниже некоторые выводы этих исследований, важные для нашего анализа, и проиллюстрируем их примерами.

Наличие существенных различий между WoS, Scopus и Google Scholar по охвату научных дисциплин было подтверждено во многих исследованиях (см., напр.: [7], см. также: [1; 2]); особенно «проседающими» для первых двух баз являются области социальных и гуманитарных наук. Исследователи делают выводы, что новые базы Microsoft Academic и Dimensions способны выступать альтернативой WoS и Scopus с точки зрения публикационного охвата [8; 9; 10], однако при этом WoS и Scopus по-прежнему остаются самыми популярными источниками информации для наукометрических исследований [11].

Для целей настоящего исследования важно остановиться на исследовательском дизайне работ, посвященных сравнению различных баз данных, и определить, какие параметры обычно выступают основаниями для сравнения. Наиболее прямым методом сравнения охвата документов из разных источников данных является получение полных списков всех документов, их сопоставление и оценка размера совпадений, что сложно осуществимо в связи с объемом данных и приводит к необходимости формиро-

вания выборок [9]. С точки зрения формирования массивов данных для анализа единицами анализа могут выступать *публикации*, выборки которых формируются на основе идентичных поисковых запросов по ключевым словам или через сплошной сбор по отобранным журналам, научным дисциплинам, областям наук, группам авторов, университетам или странам (но иногда отбор происходит и в случайном порядке), а также *журналы*, индексируемые в базах, выборки которых формируются экспертным образом. Как правило, для анализа задается определенный период времени; начальной точкой часто выступает год запуска самой новой площадки, участвующей в сравнении. В связи с этим выборки часто «гетерогенно разнообразны» [9] и ограничены по объему, а сравнение осуществляется по ограниченному числу научных баз (чаще всего – двум или трем площадкам). Однако есть и довольно обширные исследования – например, сравнение WoS, Scopus и Google Scholar по 37 научным направлениям в динамике [7], систематическое сравнение трех упомянутых баз и Microsoft Academic, Dimensions и OpenCitations' COCI по 252 категориям [9] или сравнение 12 академических поисковых систем и библиографических баз данных [12]. Еще один возможный вариант формирования выборки – использование канонического набора публикаций как исходной выборки документов ("seed sample"), для которого во всех анализируемых базах находятся все цитирующие их документы, которые и становятся выборками по каждой базе [9]. Полученные различными способами массивы данных могут сравниваться для изучения покрытия баз данных по следующим параметрам.

Характеристики базы [12]:

- *тип базы и доступа* – библиографическая база, поисковая система, агрегатор; платный/бесплатный доступ; открытый /закрытый / доступный по запросу контент; владелец;
- *функциональные возможности* поиска, анализа и экспорта данных, транспарентности алгоритмов (например, обработки

запросов и ранжирования документов на странице результатов);
– *охват* – типы индексируемых документов, скорость их индексации, предметный охват, годы покрытия, доступные поля для поиска (метаданные), качество метаданных (например, соотнесение разделения по предметным областям в разных базах [8]).

Объем:
– *по публикациям* (количество отобранных документов) – например, распределение числа публикаций, найденных по запросу или через исходную выборку документов, в том числе во временной перспективе [7; 9; 10; 12], и подсчет ежегодных темпов роста, в том числе по российским публикациям [13];
– *по журналам* (количество отобранных журналов) – количество журналов, входящих в основные списки журналов (master lists) в анализируемых базах [8].

Распределение публикаций:
– *по журналам* – абсолютное количество и доли публикаций в анализируемых журналах в анализируемых базах [10; 11];
– *по типам документов* – абсолютное количество и доли публикаций типа «статья», «ревю», «глава в книге» и др. в анализируемых базах [8; 10; 13];
– *по предметным областям* – абсолютное количество, доли, среднее число, темпы роста публикаций по основным научным областям / предметным категориям [7; 9; 11] в анализируемых базах, в том числе с учетом распределения по 20 отобранным для анализа странам [8] и используемым языкам [14], а также для российских публикаций [13];
– *по странам* – абсолютное количество публикаций в анализируемых базах, их доли в общем объеме научных исследований в мире, годовые темпы роста для 20 отобранных стан [8] или стран, лидирующих по этим показателям [11; 13];
– *по языку* – доли публикаций на разных языках в общем объеме публикаций в анализируемых базах с учетом их исследовательских направлений или во временной перспективе [14], в том числе русскоязычных и нерусскоязычных [13];
– *по количеству полученных внутри базы цитирований*;
– *по типу коллаборации* – долям публикаций, написанных в коллаборации, в том числе во временной перспективе [13].

Распределение авторов:

– *по полученным внутри базы цитированиям* (и вариациям этой метрики) – среднее число полученных авторами цитирований по различным дисциплинам в анализируемых базах [7];
– *по специализированным индексам оценки исследователей, таким как Индекс Хирша (и вариациям этой метрики),* – например, средние h- и hIa-индексы для исследователей, сгруппированных по научным дисциплинам в анализируемых базах [7];
– *по странам*.

Распределение журналов:

– *по полученным внутри базы цитированиям* – количество цитирований журналов и доля от максимального числа цитирований в анализируемых базах [10];
– *по дисциплинам*.

Пересечение между массивами:

– *по публикациям* – доли пересечений между массивами публикаций, найденные через сопоставление по DOI или названию и авторам (например, с помощью расстояния Дамерау-Левенштайна) [9];
– *по журналам* – доли пересечений между списками журналов через сопоставление по названиям / аббревиатурам / ID журналов [8].

В контексте изучения национальных корпусов научной литературы важно остановиться на вопросе представленности неанглоязычных авторов в международных базах. WoS CC включает национальные индексы цитирований для Китая, Латинской Америки и Южной Африки, Кореи, России и арабского региона, которые формируются в кооперации с представителями этих стран и потенциально должны приводить к большей представленности национальных корпусов. Большое значение для представленности работ имеют официальные языки публикаций, принятые на площадках. Интересно, что начало использования русского языка как публикационного в Scopus привело к росту доли работ российских авторов с 4,8 до 14,8% в 2006–2016 гг., что во многом объясняет экспоненциальный рост числа отечественных публикаций, наблюдаемый в эти годы [13]. Вместе с тем исследования, посвящен-

ные изучению вопросов покрытия WoS и Scopus по различным странам и языкам публикации (см., напр., обзор в работе: [14]), показали чрезмерную представленность в этих базах журналов на английском языке и публикаций из англоязычных стран (более 92% в 2018 г.). Вторым языком в 2018 г. был назван китайский, третьим – испанский. Однако даже в случае представленности языка в базе библиографические данные неанглоязычных авторов могут содержать ошибки. Так, например, исследователи обнаружили, что для испанского языка около 50% имен авторов имеют несколько вариаций написания даже внутри одной и той же международной базы [15]. Если же речь идет о разных базах, то представленные в них библиографические описания одних и тех же статей тоже не всегда консистентны (например, это показывает анализ публикаций ученых Южной Африки [16]).

Крупнейшей базой научных публикаций в России, а также научной периодики на русском языке в мире является научная электронная библиотека eLibrary. Как было сказано выше, эта платформа интегрирована с Российским индексом научного цитирования (базой РИНЦ), куда входят публикации в индексируемых (соответствующих определенным критериям качества) российских научных журналах и непериодических изданиях. С 2016 г. часть публикаций из РИНЦ (порядка 600–700 лучших российских журналов по всем научным направлениям за последние 10 лет) индексируются в WoS в виде отдельной базы Russian Science Citation Index (RSCI). Вместе с публикациями российских авторов, индексируемыми в Scopus и WoS, база RSCI составляет «ядро РИНЦ» – более узкую базу по отношению к РИНЦ. Сама площадка eLibrary при этом шире, чем РИНЦ, так как содержит неиндексируемые в этой базе публикации из изданий, имеющих заключенный с площадкой договор.

Несмотря на наличие отдельной национальной базы научных публикаций, eLibrary в какой-либо из трех вариаций (от максимально полной до ядра РИНЦ) редко выступает источником

для сравнения с другими площадками. Например, в исследовании коллекций публикаций российских авторов [13] сравниваются коллекции публикаций из баз WoS CC и Scopus, но отдельно подчеркивается, что базы RSCI и РИНЦ в анализе не участвуют. Сравнение RSCI с международными базами проводилось аналитиками eLibrary [17] посредством анализа массивов публикаций российских авторов в «квартильных» журналах WoS CC и Scopus, в базе новых источников ESCI (Emerging Sources Citation Index) из ядра WoS и базы RSCI. По всем массивам было подсчитано число публикаций; на основании информации об индексации журналов в различных базах были найдены пересечения между массивами; для RSCI и ESCI и для выделенных в WoS и Scopus групп по квартилям были подсчитаны средние показатели цитирования внутри базы. Проведенный анализ позволил говорить только о частичном пересечении коллекций журналов в трех базах и об оригинальности значительной части контента базы RSCI и большом вкладе в ядро РИНЦ.

Отдельно базы РИНЦ и RSCI (также встречается название RSCI-C – от названия компании Clarivate [13]) выступают источниками данных в библиометрических исследованиях, ориентированных на изучение российского научного поля (см., напр.: [18; 19]). Однако методология сбора данных для баз РИНЦ и RSCI детально не описывается. Отдельным методологическим затруднением может быть то, что на английский язык «Российский индекс научного цитирования» переводят не только как “Russian Index of Science Citation” (RISC), но и как “Russian Science Citation Index” (RSCI), что повторяет название в базе WoS – хотя, как видно из обзора, это хоть и смежные, но разные базы.

Рассмотренные обзорные работы, сравнивающие разные базы на уровне стран (см.: [8]), показывают, что использование информации из разных баз данных может привести к различным результатам библиометрической оценки деятельности национальных научных коллективов. Так, исследователи пришли к выводу,

что и база WoS, и особенно Scopus должны с осторожностью применяться в качестве единственных инструментов измерения результативности неанглоязычных исследователей, в том числе российских [13]. В связи с различиями в покрытии исследователи рекомендуют использовать *несколько* баз данных для формирования наиболее полного массива исследования. В случае изучения развития науки в конкретных странах рекомендуется формировать массив данных из публикаций в международных и национальных базах научного цитирования. Первым шагом для проведения таких исследований является сравнение международных и национальных баз друг с другом. Такой анализ и предпринимается в данной статье.

Методология и данные исследования

Методология

На основе рассмотренных параметров сравнения баз данных научных публикаций, а также предварительного анализа рассматриваемых массивов были сформулированы основания и параметры для сравнения баз WoS и eLibrary (табл. 1), которые касаются: 1) получения доступа к данным, 2) организации данных в базах, 3) количественных характеристик (объем, динамика и т.д.) и 4) содержательных характеристик указанных данных. Для описания особенностей доступа к данным и организации данных в базах использовался описательный анализ доступной информации. Согласно стратегиям исследований, рассмотренных в обзоре, дизайн исследования для более детального анализа подразумевал сравнение аналогичных массивов данных, выгруженных из каждой базы. Два массива рассматривались по определенным параметрам с помощью статистического анализа, затем на основе данных из массивов было построено несколько базовых библиометрических сетей, которые также рассматривались в соотношении друг с другом по ряду рассчитанных параметров.

Анализ указанных параметров позволяет найти пересечения между массивами с точки зрения присутствующих в них единиц анализа – публикаций, журналов, авторов, ключевых слов – и оценить размер множеств, находящихся на пересечении и

Таблица 1

ОСНОВАНИЯ, ПАРАМЕТРЫ И СПОСОБЫ АНАЛИЗА ДЛЯ СРАВНЕНИЯ БАЗ ДАННЫХ

Основания	Параметры	Способы анализа
Получение доступа к данным	Особенности и возможности сбора данных исследователем	Описательный анализ
Организация данных в базах	Формат и структура получаемых массивов, в том числе количество используемых в библиографических описаниях метаданных. Возможности предобработки данных существующим программным обеспечением. Возможности построения файлов для сетевого библиометрического анализа	Описательный анализ
Количественные характеристики данных	Объем массивов (число публикаций). Динамика числа публикаций во времени. Объем пропущенных значений по метаданным в библиографических описаниях публикаций. Количество уникальных библиометрических единиц (публикаций, авторов, журналов, ключевых слов)	Статистический анализ

Окончание табл. 1

Основания	Параметры	Способы анализа
Содержательные характеристики данных	Распределение уникальных авторов и ключевых слов по частоте встречаемости в публикациях в массиве. Распределение соавторов по авторам в массиве. Наиболее частотные журналы, ключевые слова. Наиболее популярные авторы по числу работ и числу соавторов	Статистический и сетевой анализ двумодальных сетей работ и авторов WA, работ и ключевых слов WK, одномодальной сети коллабораций Co.

при объединении массивов. Помимо этого, сравнение результатов анализа с точки зрения содержания также позволяет сделать выводы о том, насколько похожие результаты дает использование двух баз данных.

Методология анализа данных для количественной оценки по выделенным основаниям подразумевала использование различных инструментов статистического и сетевого анализа. Процедура анализа и использованное программное обеспечение представлены в тексте ниже при сравнении показателей.

С точки зрения используемых инструментов анализа подсчет общей статистики по набору данных WoS проводился с помощью пакета Bibliometrix в R и его приложения Biblioshiny.

Данные

В статье сравниваются публикации российских социологов на площадках WoS Core Collection и eLibrary. Единицами анализа являются научные статьи в научных журналах в области социологии. Оба массива данных включают публикации за 2010–2021 гг. Для сбора данных на двух площадках использовались идентичные

стратегии. Ниже представлена информация о деталях сбора, размере массивов и формате итоговых данных.

Стратегия сбора данных и размеры массивов. Анализируемый массив данных собран в рамках проекта, выполняемого в рамках гранта РНФ[1]. При сборе данных из eLibrary работа проводилась совместно с сотрудниками ООО «Научная электронная библиотека», осуществляющими поддержку этой базы. Поиск работ проводился по всем научным журналам, представленным на сайте eLibrary (имеющим заключенный договор). Из всех работ, относящихся по ГРНТИ к рубрике «Социология», были отфильтрованы публикации типа «научная статья», где по крайней мере одним из авторов является российский ученый (в поле «страна» указано «Россия»). По данному запросу был составлен список из 75 232 уникальных идентификаторов публикаций, по которым затем была собрана полная библиографическая информация. В сравниваемый массив данных eLibrary вошло 75 232 публикации.

Анализируемый массив данных WoS CC является подмножеством из набора данных, собранных в рамках проекта по изучению российской науки на основе всех российских публикаций, представленных в WoS CC[2] (1 383 996 библиографических записей о российских публикациях по всем наукам за период с 1992 г. до мая 2022 г.). Стратегия сбора данных этого массива подразумевала использование базы Core Collection. Были отобраны и выгружены все публикации российских авторов (поле CU = “Russia”). Затем на основе категории (“research area”), к которой относится публикация, было выделено подмножество по социологии (поле

[1] Проект «Паттерны коллаборации в российском социологическом сообществе: структура научных школ и возможные точки роста» выполнен в рамках гранта Российского научного фонда в 2021–2023 гг. под руководством Д.В. Мальцевой.

[2] Проект осуществляется совместно Д. Фиалой, отвечающим за сбор и исследовательский анализ данных, и коллективом МЛ ПСА под руководством Д.В. Мальцевой, отвечающим за сетевой анализ массива.

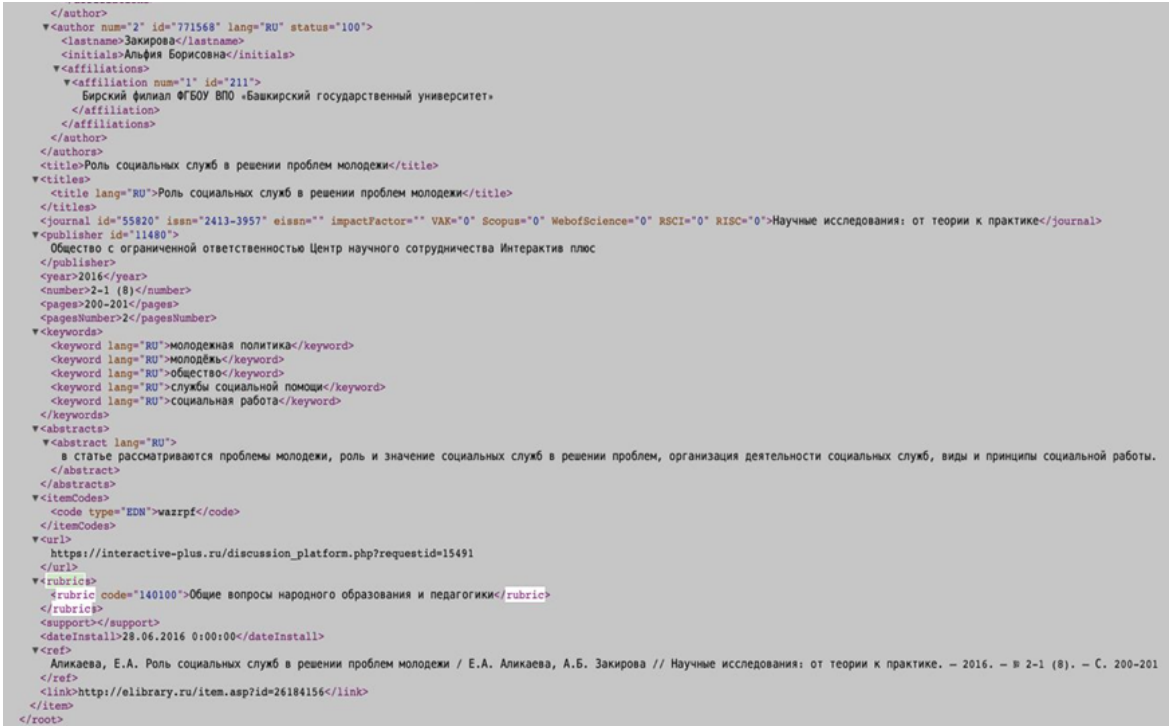
SC = “Sociology”). Первоначально массив состоял из 7915 публикаций, но для целей настоящего анализа он был ограничен по типу публикаций (поле DT = “Article”) и временному периоду (2010–2021 гг.), что дало 3559 научных публикаций, которые и вошли в сравниваемый массив данных WoS CC.

Выгрузка данных. База данных eLibrary не предоставляет функциональных возможностей для выгрузки библиографических описаний публикаций с сайта или публичного адреса API-сервиса для автоматизированного парсинга (сбора и структурирования информации) данных. Доступ к закрытому API-сервису возможен при наличии договора с ООО «НЭБ», который был заключен в рамках исследования. Для сбора данных, а также их предобработки и построения сетевых файлов использовался разработанный авторами статьи методологический подход [20], реализованный в программе Bib-eLib[1]. Используя полученный список из идентификаторов публикаций через соответствующий сервис API с помощью специально написанного парсера, делались запросы на информацию по каждой публикации. Выдача данных представляет собой XML-страницу структурированного вида с идентифицированными полями (рис. 1).

Выгрузка данных осуществлялась из нужных полей, а затем записывалась в единый файл формата csv. Данные собраны в октябре 2022 г.

Платформа WoS дает возможность выгрузки библиографических описаний отобранных работ в различных форматах (RIS, Excel, обычный текстовый файл – plain text) для всех зарегистрированных пользователей, чей доступ осуществляется через

[1] Программа для ЭВМ Bib-eLib для сбора и обработки библиографических данных на русском языке из электронной библиотеки eLibrary на языке программирования Python зарегистрирована в виде РИД (свидетельство о государственной регистрации программы для ЭВМ № 2023684182, регистрация в реестре программ для ЭВМ 14.11.2023) и доступна в репозитории платформы GitHub по ссылке: https://github.com/Daria-Maltseva/Collaboration (дата обращения: 08.04.2024).



```
    </author>
  ▼<author num="2" id="771568" lang="RU" status="100">
      <lastname>Закирова</lastname>
      <initials>Альфия Борисовна</initials>
    ▼<affiliations>
      ▼<affiliation num="1" id="211">
          Бирский филиал ФГБОУ ВПО «Башкирский государственный университет»
        </affiliation>
      </affiliations>
    </author>
  </authors>
  <title>Роль социальных служб в решении проблем молодежи</title>
▼<titles>
    <title lang="RU">Роль социальных служб в решении проблем молодежи</title>
  </titles>
  <journal id="55820" issn="2413-3957" eissn="" impactFactor="" VAK="0" Scopus="0" WebofScience="0" RSCI="0" RISC="0">Научные исследования: от теории к практике</journal>
▼<publisher id="11480">
    Общество с ограниченной ответственностью Центр научного сотрудничества Интерактив плюс
  </publisher>
  <year>2016</year>
  <number>2-1 (8)</number>
  <pages>200-201</pages>
  <pagesNumber>2</pagesNumber>
▼<keywords>
    <keyword lang="RU">молодежная политика</keyword>
    <keyword lang="RU">молодёжь</keyword>
    <keyword lang="RU">общество</keyword>
    <keyword lang="RU">службы социальной помощи</keyword>
    <keyword lang="RU">социальная работа</keyword>
  </keywords>
▼<abstracts>
  ▼<abstract lang="RU">
      в статье рассматриваются проблемы молодежи, роль и значение социальных служб в решении проблем, организация деятельности социальных служб, виды и принципы социальной работы.
    </abstract>
  </abstracts>
▼<itemCodes>
    <code type="EDN">wazrpf</code>
  </itemCodes>
▼<url>
    https://interactive-plus.ru/discussion_platform.php?requestid=15491
  </url>
▼<rubrics>
    <rubric code="140100">Общие вопросы народного образования и педагогики</rubric>
  </rubrics>
  <support></support>
  <dateInstall>28.06.2016 0:00:00</dateInstall>
▼<ref>
    Аликаева, Е.А. Роль социальных служб в решении проблем молодежи / Е.А. Аликаева, А.Б. Закирова // Научные исследования: от теории к практике. – 2016. – № 2-1 (8). – С. 200-201
  </ref>
  <link>http://elibrary.ru/item.asp?id=26184156</link>
</item>
</root>
```

Рис. 1. Пример библиографического описания публикаций в eLibrary (XML-страница)

Document1 * ×

```
PT J
AU Toshchenko, ZT
AF Toshchenko, ZT
TI Sociology of life as a concept to study social reality
SO SOTSIOLOGICHESKIE ISSLEDOVANIYA
LA Russian
DT Article
AB "Sociology of life as a concept to study social reality" begins with an analysis of reasons for absence of sociological schools in contemp
C1 Russian Acad Sci, Moscow, Russia.
RP Toshchenko, ZT (corresponding author), Russian Acad Sci, Moscow, Russia.
CR ANDRUSHENKO VP, 1996, SOTSIOLOGIYA NAUKA O
   BARAZGOVA ES, 1997, SOTSIS
   ...
   YADOV VA, 1998, STRATEGIYA SOTSIOLOG
NR 23
TC 5
Z9 12
U1 0
U2 3
PU IZDATELSTVO NAUKA
PI MOSCOW
PA PROFSOYUZNAYA UL 90, MOSCOW, 117864, RUSSIA
SN 0132-1625
J9 SOTSIOL ISSLED+
JI Sotsiologicheskie Issled.
PY 2000
IS 2
BP 3
EP +
PG 11
WC Sociology
WE Social Science Citation Index (SSCI)
SC Sociology
GA 292NH
UT WOS:000085798200001
DA 2022-05-09
ER
```

Рис. 2. Пример библиографического описания публикаций в WoS (текстовый файл)

организационную подписку. Используемый формат представляет собой текстовый файл структурированного вида с идентифицированными полями (рис. 2).

В зависимости от того, собирается или нет информация о цитируемой литературе (поле “CR”), за одну итерацию можно загрузить до 500 или 1000 библиографических описаний в едином файле. Для сбора данных был написан программный код на Python, который позволял итеративно обращаться к базе и последовательно собирать файлы с описаниями в формате .txt, которые затем были автоматически собраны в единый файл. Данные собраны в мае 2022 г.

Формат и структура данных. Собранный массив данных eLibrary представлен в виде таблицы в формате .csv, в которой приведена информация по всем полям, доступным к выгрузке, для всех 75 232 публикаций. Данные WoS CC представлены в виде единого текстового файла в формате .txt с полным библиографическим описанием 3559 публикаций, включая пристатейные списки литературы. В обоих массивах содержится информация по таким библиографическим единицам как публикация, автор(ы) и журнал. Набор метаданных, которыми описываются библиографические единицы в каждом массиве, приведен в табл. 2.

Таблица 2

МЕТАДАННЫЕ БИБЛИОГРАФИЧЕСКИХ ОПИСАНИЙ
В WoS И eLibrary

№	Параметр	WoS	eLibrary
Публикация			
1	ID публикации	+	+
2	Название на русском языке	-	+
3	Название на английском языке	+	+
4	DOI публикации	+	+
5	Дата публикации (год)	+	+
6	Предметная область	+	+

Окончание табл. 2

№	Параметр	WoS	eLibrary
7	Язык публикации	+	-
8	Тип публикации	+	-
9	Количество страниц (начальная и конечная страницы)	+	+
10	Число цитирований	+	+
11	Число использований (доступ, скачивание)	+	-
12	Аннотация на русском языке	-	+
13	Аннотация на английском языке	+	+
14	Ключевые слова на русском языке	-	+
15	Ключевые слова на английском языке	+	+
16	Информация о финансировании	-	+
17	Гиперссылка на статью в базе	-	+
18	Библиографическое описание	-	+
19	Список цитируемой литературы	+	
20	Количество процитированной литературы	+	
Автор(ы)			
1	Фамилия и инициалы на русском языке	-	+
2	Фамилия и инициалы на английском языке	+	+
3	ID автора	-	+
4	Название аффилиации автора	+	+
5	ID аффилиации автора	-	+
6	Местоположение (страна, город)	+	-
Журнал			
1	Название журнала	+	+
2	ID журнала	-	+
3	ISSN/e-ISSN	+	+
4	Импакт-фактор	-	+
5	Включенность в другие базы данных	+-	+
6	Выпуск	+	+
7	Номер	+	+
8	Название издательства	+	+
9	ID издательства	-	+
10	Адрес издательства (город и почтовый адрес)	+	-

В целом можно видеть, что базы похожи по используемым ими метаданным. Очевидное отличие базы WoS CC заключается в том, что в ней не представлена информация на русском языке (название, аннотация, ключевые слова, имя автора). В этой базе также отсутствует информация об ID авторов и их организаций, ID журналов и издательств – хотя она может дать важные идентифицирующие признаки при решении проблемы дизамбигуации единиц анализа (но нужно отметить, что для журналов WoS помимо полного названия приводит и два вида его стандартизированной аббревиатуры, что может быть использовано для обозначенной цели). В этом смысле наличие ID в описаниях eLibrary выгодно отличает эту базу и предоставляет исследователям дополнительные аналитические возможности. В данных eLibrary также указываются базы, в которые входит публикация (РИНЦ, RSCI, WoS, Scopus и ВАК); в WoS предоставляется информация только о базе внутри ядра CC, к которой принадлежит публикация.

Главным выгодным отличием базы WoS CC является наличие информации о цитируемой литературе (поле “CR”), что позволяет проводить определенные виды анализа – изучать сети цитирований, социтирований, библиографического сочленения между различными библиографическими единицами (авторами, публикациями, журналами и т.д.). В обеих базах подсчитывается также число цитирований, полученных внутри данной базы и других баз. Помимо цитирования, в WoS есть также показатель по использованию публикации другими авторами за определенные периоды времени (доступу к тексту и загрузке), что также показывает внешний интерес к научной работе.

Предобработка данных. Поскольку библиографические описания публикаций не всегда содержат полную информацию или приведенная информация может содержать ошибки [20], для повышения качества дальнейшего анализа перед построением сетевых файлов возникает задача по предобработке данных, т.е. устранению пропущенных значений и приведению единиц

анализа к единому виду. Предложенная в авторской методологии [18] логика предобработки данных массива eLibrary отчасти следует логике работы с данными WoS, поэтому вначале рассмотрим предобработку данных, реализованную для массива данных WoS.

Предварительный исследовательский анализ данных массива WoS в программе Biblioshiny (табл. 3) показал, что значительная часть данных отсутствует в полях аннотаций и ключевых слов (поле “DE”, Keywords – около 21%) и в полях с дополнительными ключевыми словами (поле “ID”, Keywords Plus и DOI – 50 и 79% соответственно).

Таблица 3

ОЦЕНКА ПОЛНОТЫ БИБЛИОГРАФИЧЕСКИХ МЕТАДАННЫХ В МАССИВАХ WoS И eLibrary: ПРОПУЩЕННЫЕ ЗНАЧЕНИЯ

Пропущенные элементы метаданных	WoS		eLibrary	
	число	доля, %	число	доля, %
Публикация				
Название – английский язык	0	0	67 048	89,1
Название – русский язык			0	0
Аннотация – английский язык	737	20,71	27 739	36,9
Аннотация – русский язык			27 751	36,9
Ключевые слова – английский язык	774	21,75	27 600	36,7
Ключевые слова – русский язык			11 568	15,4
Дополнительные ключевые слова (поле Keywords Plus)	2806	78,84		
DOI	1780	50,01	62 247	82,7
Год публикации	0	0	0	0
Количество цитирований работы	0	0	0	0
Количество цитируемых статей	0	0		
Цитируемые источники	235	6,60		
Число страниц	0	0	0	0
Информация о финансировании	2951	82,9	73 447	97,6

Окончание табл. 3

Пропущенные элементы метаданных	WoS		eLibrary	
	число	доля, %	число	доля, %
Автор(ы)				
Автор (фамилия и имя) – английский язык	0	0	35	0,05
Автор (фамилия и имя) – русский язык			1 539	2,05
Название аффилиации автора на русском	-	-	4609	6,1
Название аффилиации автора на английском	2	0,06	37 891	50,4
Журнал				
Журнал (название)	0	0	0	0
Журнал (ID)			0	0
Информация об издателе	0	0	1223	1,6

Примечание. Знак "-" означает, что данное поле в базе не представлено. В ключевых словах на английском языке для WoS указаны данные из поля "DE". Более темным цветом выделены ячейки с более высокими долями пропущенных данных.

Предобработка данных для массива WoS проводилась с помощью программы WoS2Pajek [21], используемой для предобработки сетевых данных и построения сетевых файлов. С помощью этой программы была произведена чистка исходного файла с библиографическими описаниями – автоматическая идентификация и удаление дублей публикаций и лишних символов в файле. Для формирования массива ключевых слов программа берет информацию из полей "DE" – Keywords, "ID" – Keywords Plus, а также из названий статей и аннотаций – полей "TI" – Title и "AB" – Abstract, что решает проблему неполного покрытия некоторых из этих полей (обозначенную программой Biblioshiny). Программа проводит нормализацию и приводит к единому виду ключевые слова, используя словари для английского языка.

Поскольку программа WoS2Pajek ориентирована на использование информации о цитировании работ (кратких описаний цитируемых публикаций в поле “CR”), фокус делается на обработке библиографических описаний. Работы, указанные в поле “CR”, записаны в формате: AU + ’, ’ + PY + ’, ’ + SO[:20] + ’, V’ + VL + ’, P’ + BP (автор, год публикации, до 20 символов источника публикации / журнала, выпуск, начальная страница), например: TOSHCHENKO ZT, 2000, SOTSIOL ISSLED, V23, P123. Изначально такой подход использовался для повышения точности данных при внесении информации по единому формату. Но так как по факту одна работа может иметь отличающиеся наименования, для повышения точности программа WoS2Pajek использует короткие имена, записываемые в формате: LastNm[:8] + ’ ’ + FirstNm[0] + ’(’ + PY + ’)’ + VL + ’:’ + BP (8 символов фамилии, первая буква имени, год публикации, выпуск журнала, начальная страница), например: TOSHCHEN_Z(2000)23:123. Та же самая процедура создания коротких имен осуществляется и для работ, имеющих полные библиографические описания («хитов»). Для решения проблемы дизамбигуации имена авторов записываются по форме: LastNm[:8] + ’ ’ + FirstNm[0] (8 символов фамилии, первая буква имени), например: TOSHCHEN_Z. Безусловно, при таком подходе могут возникать проблемы «склейки» имен авторов, однако эти проблемы разрешаются путем проверки результатов, получаемых для наиболее важных единиц анализа[1]. Чистка данных, как правило, осуществляется итеративно – при нахождении проблем правки либо вносятся в исходный файл, который снова проходит через программу WoS2Pajek, либо устраняются алгоритмическим образом в программе Pajek.

Предварительный исследовательский анализ массива данных eLibrary показал, что некоторые важные для анализа параметры

[1] Методология следует так называемому статистическому подходу, согласно которому даже при некоторой неконсистентности в данных общие тренды и важные единицы анализа могут проявиться при анализе.

метаданных имеют отсутствующие значения (0 или “none”): из 37 302 уникальных авторов в начальном массиве только у 19 739 авторов имелись РИНЦ ID (хотя наличие именно этой информации рассматривалось как преимущество базы). В ходе предобработки данных eLibrary с целью дизамбигуации имен авторов собранная база трансформировалась: вначале была проведена нормализация имен авторов и их аффилиаций, а затем созданы универсальные ID для авторов в формате: eLibrary_ID + FirstNm[:2] + LastNm[:8] + Affiliation_ID (ID автора в РИНЦ, инициалы, 8 символов фамилии, ID организации автора), например: 1382_ZHT_Toschenk_5350 (подробнее см.: [20]). В результате предобработки количество уникальных названий аффилиаций сократилось на 39,5% за счет нормализации и приведения к единому виду описаний аффилиаций и создания универсальных описаний для аффилиаций с единым ID; количество уникальных ID организаций сократилось на 1,5% – за счет удаления некорректно заполненных ID, а количество уникальных авторов увеличилось на 95% – до 37 790 – за счет идентификации авторов, не имеющих РИНЦ ID [20].

Построение сетевых файлов. Используемый авторами подход к работе с данными WoS CC подразумевает использование программы WoS2Pajek [21] для трансформации массива в коллекцию связанных сетей, в частности (используемых для анализа) двумодальных сетей «Работа – Автор» (“Work – Author”) **WA**, «Работа – Ключевое слово» (“Work – Keyword”) **WK**, «Работа – Журнал» (“Work – Journal”) **WJ** (где в первом наборе указаны все публикации, во втором – авторы, ключевые слова или журналы, а далее фиксируются связи между ними). Также создается файл с информацией о годах публикаций работ (Year.clu), на основании которого сети можно разделить на периоды для изучения в динамике и файл с разделением работ – на источники с полным библиографическим описанием («хиты») и только цитируемую литературу (DC.clu).

По этой же логике после формирования массива в eLibrary из соответствующих полей с помощью специально написанного на Python программного кода были выгружены данные для построения двумодальных сетей «Работа – Автор» **WA**, «Работа – Ключевое слово» **WK** и «Работа – Журнал» (“Work – Journal”) **WJ**. Файлы сохранены в формате .net и доступны для дальнейшего анализа в программе Pajek. Был сформирован файл с информацией о годах публикаций отобранных работ[1].

Хотя напрямую сравнить два использованных подхода и инструмента для предобработки и построения сетевых файлов затруднительно, можно сделать выводы о времени, необходимом для работы в каждом случае. Ввиду автоматизированности процесса процедура предобработки данных и построения сетей с помощью программы WoS2Pajek занимает считанные минуты (построение сетевых файлов из начального подмассива заняло 2 мин. 34 сек.). Использование разработанного в рамках проекта программного кода для предобработки и подготовки сетевых файлов из массива eLibrary, зарегистрированного в виде ЭВМ, на данный момент требует гораздо больше времени ввиду необходимости ручной проверки данных на некоторых этапах.

Сравнительный анализ массивов данных

Полнота библиографических метаданных. Безусловно, не все библиографические описания в базах WoS и eLibrary содержат все возможные метаданные, обозначенные в табл. 2. Вместе с тем полнота представления метаданных библиографических описаний оказывает значительное влияние на качество анализа, поэтому важно оценить объем пропущенных данных в массивах. Табл. 3

[1] Файлы создавались для проводимого анализа, но могут быть построены и другие двумодальные сети и дополнительные файлы с атрибутами узлов (количество страниц, цитирование), которые можно использовать для решения разных исследовательских вопросов.

представляет данные для оценки полноты библиометрических описаний публикаций в двух рассматриваемых массивах – количество и долю пропущенных значений по отобранным метаданным[1] (уже после проведенной предобработки данных). Для удобства метаданные сгруппированы по типам библиометрических единиц анализа, к которым они относятся (как в табл. 2). Обратим внимание, что данные подсчитаны для публикаций: отсутствующие значения по именам авторов показывают количество статей, в которых не имеется хотя бы одного имени автора на русском и английском языках; данные в названиях аффилиаций подсчитывают число случаев, когда при наличии авторов хотя бы у одного из них отсутствует название аффилиации[2].

Согласно оценке программы Biblioshiny, доли пропущенных значений от 20 до 50% говорят о слабой, а более 50% – о критической представленности данных в поле библиографического описания и не рекомендуются программой для использования в анализе. Как видим, для массива WoS это (от наибольшей доли пропущенных значений к наименьшей) информация о финансировании, дополнительным ключевым словам, DOI (критично), ключевым словам и аннотации публикации на английском языке (слабо). Для массива eLibrary это информация о финансировании, названию на английском языке, DOI, названию аффилиации автора на русском (критично), аннотации на русском и английском, ключевым словам на английском языке (слабо). Ключевые слова на русском языке отсутствуют в 15% публикаций, однако этот показатель считается «проходным».

Основная информация по массивам. В табл. 4 собрана основная информация по числу различных единиц анализа в рассматриваемых массивах.

---

[1] За основу структуры взята таблица, формируемая в программе Biblioshiny при загрузке массива данных.

[2] Если автора нет (none и по столбцу с английской фамилией, и по столбцу с русской), наличие аффилиации не проверяется.

Таблица 4

ЧИСЛО ЕДИНИЦ АНАЛИЗА В МАССИВАХ WoS И eLibrary

Единица анализа / Количество	eLibrary	WoS	Доля WoS к РИНЦ, %
Публикации	75 232	**3559**	4,7
Журналы	3910	**109**	2,8
Авторы	37 790	3238	8,6
Ключевые слова на английском	91 109	6750	7,4

Примечание. Расчет по WoS приведен по программе WoS2Pajek.

Обратим внимание, что для сравнения в eLibrary взяты данные по числу ключевых слов на английском языке (аналогичный показатель на русском языке составляет 100 594); ключевые слова были взяты в формате, приведенном авторами, и не подвергались обработке (чем можно объяснить их большое количество – из-за наличия множества уникальных слов). По массиву WoS подсчет приведен по данным программы WoS2Pajek[1]. Для информации подсчитано соотношение единиц из массива WoS к массиву eLibrary, показывающее значительное превосходство данных eLibrary по объему.

Динамика количества **публикаций** в обеих базах за рассматриваемый период показана на рис. 3.

Распределение абсолютного числа публикаций (рис. 3) в eLibrary показывает плавный рост и достижение максимума в 2016 г. и следующее за ним снижение. Количество российских социологических публикаций в WoS достигает максимума в 2019 г., однако далее снижается незначительно. Чтобы увидеть

[1] В связи с имплементированными алгоритмами предобработки Biblioshiny выделяет несколько иное число авторов (3554) и ключевых слов (12 215). Данные по ключевым словам, полученные WoS2Pajek, следует рассматривать как более валидные – ввиду более точного подсчета (для Biblioshiny число рассчитано как сумма по полям ID и DE, нет возможности учета пересечений) и заложенных в программу алгоритмов нормализации ключевых слов.

общие тренды, данные были подсчитаны кумулятивно и затем нормированы в диапазоне от 0 до 1 (значение за каждый год разделено на сумму публикаций) (рис. 4).

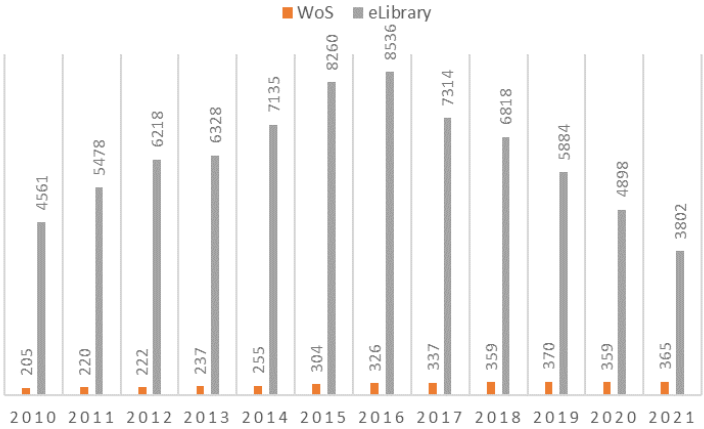


Рис 3. Динамика количества публикаций в базах WoS и eLibrary: абсолютное число публикаций

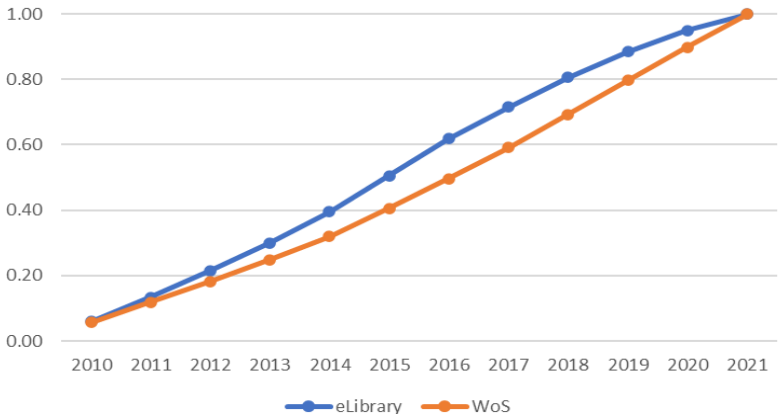


Рис 4. Динамика количества публикаций в базах WoS и eLibrary: число публикаций, нормированное на кумулятивной шкале

В такой репрезентации лучше видно, что относительные доли числа публикаций в eLibrary с 2014–2015 гг. были выше, чем в WoS. Однако если средний годовой прирост публикаций[1] в eLibrary на 2021 г. составляет -0,6%, аналогичный показатель для WoS составляет 5,5%, что говорит о более динамичном увеличении числа публикаций в этой базе.

Анализ пересечений между базами на основе статистики

Интересной находкой исследования стало то, что для каждой публикации в массиве eLibrary содержится информация о том, в каких наукометрических базах (РИНЦ, RSCI, WoS, Scopus, ВАК) она проиндексирована, что дает возможность посмотреть на распределение и пересечение публикаций в разных базах на основе информации от eLibrary. Эта информация напрямую не относится к предмету данной статьи, однако приведена в Приложении. Проведенный анализ подтвердил, что не все статьи, вошедшие в массив eLibrary (75 232 публикации), входят в базу РИНЦ. Также анализ показывает, что база РИНЦ в значительной степени пересекается с другими, меньшими по размеру базами библиографических данных, включая базу WoS. Это дает основания к проверке пересечений между двумя базами, проводимой ниже. В данном разделе сравниваются массивы данных WoS и eLibrary по публикациям и авторам, включенным в два анализируемых массива данных.

Сопоставление публикаций. Для сопоставления публикаций, входящих в базы WoS (3559) и eLibrary (75 232), было использовано несколько подходов: мы последовательно сопоставляли мас-

[1] Рассчитанный путем деления разницы между количеством публикаций за каждую пару лет (n, $n + 1$) на значение первого года (n), и последующий расчет среднего всех полученных значений за каждый год в анализируемый период времени (2010–2021 гг.).

сивы данных по: 1) DOI; 2) названию публикаций на английском языке; 3) сгенерированной комбинации из последовательности авторов и года написания статьи.

Поиск совпадающих DOI статей позволил сопоставить 655 публикаций. Ситуацию осложняло отсутствие DOI у 50% статей в WoS и у 83% в eLibrary. Далее сопоставление статей было продолжено путем сравнения их названий на английском для оставшихся неидентифицированными публикаций (обратим внимание на высокую долю пропущенных значений). Названия были предварительно предобработаны (убраны знаки препинания и цифры, все символы приведены к нижнему регистру, убраны стоп-слова: of, in, at, is; артикли) и по точным совпадениям удалось получить еще 32 совпавшие статьи. Далее мы приступили к поиску совпадений по комбинации авторов и года написания статьи для оставшихся неидентифицированными массивов данных. Например, если Михаил Соколов (*SOKOLOV MM*) и Кирилл Титаев (*TITAEV KD*) написали статью в 2014 г., их статье была присвоена строка "S*OKOLOV MM;TITAEV KD_2014*". Такие «простые ID» мы сгенерировали для всех статей в WoS и eLibrary и искали совпадения по ним. Отметим, что для повышения точности оценки и избежания ложноположительных совпадений поиск совпадений проводился только для статей, чье «простоеID» встречалось в базе данных только один раз. В противном случае совпадение будет неточным: один и тот же социолог или группа исследователей могут написать несколько статей в один и тот же год, и мы не сможем точно сопоставить, например, одну публикацию Ж.Т. Тощенко в 2012 г. в WoS с какой-то из шести публикаций Тощенко в 2012 г. в eLibrary. По результатам этого поиска нашлось еще 358 статей.

Несмотря на то, что комбинация выбранных способов поиска идентичных статей неидеальна, она позволяет получить примерную оценку совпадения двух баз данных без разработки специфических технологических решений. В теории они могли бы включать поиск совпадающих статей по разным вариациям

автоматического перевода названия статьи с русского на английский, поиска совпадающих статей по комбинации авторов с поиском возможных расхождений в один-два символа в фамилиях и пр., однако эта разработка может стать темой отдельного исследовательского проекта. Итоговое значение совпавших статей в базе данных elibrary и WoS составляет 1013 статей – 28,5% от всех публикаций в базе данных WoS или 1,4% от всех публикаций в базе данных elibrary.

Сопоставление авторов. Для сравнения авторов, присутствующих в базах WoS и eLibrary, мы выбрали подход, в котором искали совпадения по фамилиям и инициалам авторов; таким образом, ограничением для следующих оценок стало предположение о том, что одна комбинация фамилии и инициалов принадлежит одному автору. Так, хотя в предварительно обработанных данных eLibrary были созданы новые универсальные ID, позволяющие идентифицировать разных авторов (основываясь на ID РИНЦ, инициалах, фамилии и ID аффилиациии), они бы не совпадали с потенциальными ID, которые можно было бы сконструировать на основе данных WoS. WoS содержит информацию о ResearcherID и ORCID-ID исследователей, но эти поля часто не заполнены. В нашей базе в 54,5% статей нет информации о ResearcherID ни для одного из авторов статьи; для ORCID-ID аналогичная оценка составляет 61,4%.

По этим причинам в базе данных статей российских социологов из eLibrary мы создали столбец с перечислением всех авторов, аналогичный по формату записям в базе WoS, где авторы записываются в столбце “AU” следующим образом: “*KOLESNIK NV;SHOPULATOV AN;SINYUTIN MV*”. Отметим, что предобработка фамилий (например, приведение отдельных написаний фамилии к наиболее популярному виду) не производилась, однако такие процедуры можно было бы провести, тем самым повысив точность оценки. Число имен авторов (табл. 5), сформированных таким образом, не полностью совпадает с числом имен авторов

после дизамбигуации для двух массивов (табл. 4), однако такой подход позволяет дать некоторую количественную оценку имеющимся пересечениям авторов в базе и показать, какие авторы присутствуют только в одной или в обеих базах. По полученным результатам (табл. 5), число авторов в обеих базах составило 1180 авторов, что составляет 33% от всех авторов в массиве WoS, но только 3,3% от всех авторов.

Таблица 5

ПОКАЗАТЕЛИ СОПОСТАВЛЕНИЯ АВТОРОВ
В БАЗАХ WoS И eLibrary

Показатель	Значение
Число авторов в eLibrary	35 462
Число авторов в WoS	3554
Число авторов в обеих базах	1180
Доля авторов в обеих базах относительно числа уникальных авторов в данных WoS	33%
Доля авторов в обеих базах относительно числа уникальных авторов в данных eLibrary	3,3%

Примечание. Число авторов в eLibrary подсчитано так же, как в WoS (первые 8 букв имени и инициалы после разделителя), поэтому число авторов не совпадает с числом авторов из табл. 4, к которым был применен подход по дизамбигуации имен.

Анализ пересечений между базами на основе содержательных результатов

В данном разделе проводится опосредованное, непрямое сравнение того, насколько похожими являются массивы данных WoS и eLibrary с точки зрения получаемых результатов при анализе массива и производных сетей.

Работы и авторы. Входящая центральность в двумодальной сети **WA** показывает количество работ у авторов. Распределение этого показателя для двух массивов приведено на рис. 5.

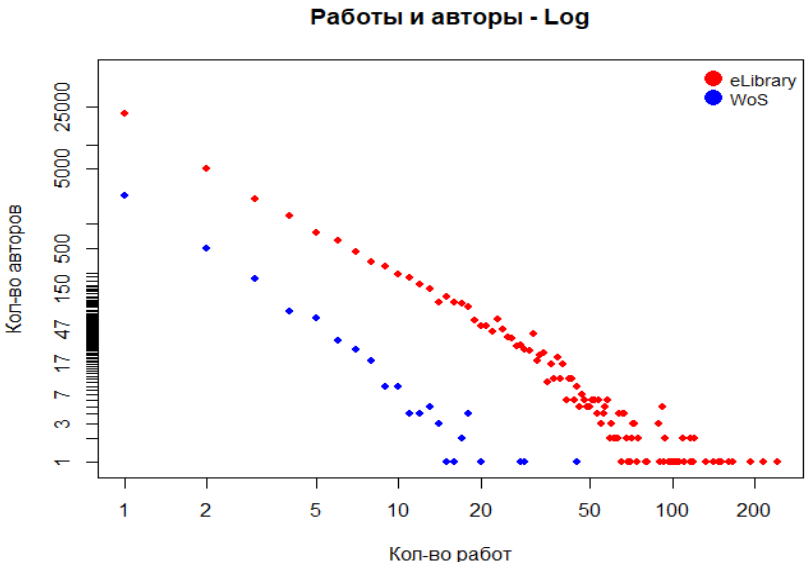


Рис. 5. Распределение количества работ по авторам в двух массивах данных (логарифмическая шкала)

Массивы данных значительно различаются по размеру – число публикаций в массиве eLibrary примерно в 20 раз больше числа публикаций в WoS – поэтому различия наблюдаются и в числе авторов. Вместе с тем распределения на рис. 5 похожи по тренду и могут следовать степенному закону, или закону Лотки, описывающему распределение продуктивности ученых[1]. Тогда как 66% авторов в массиве eLibrary и 64% в массиве WoS имеют только одну публикацию, еще 13 и 14% – две, а по 6% – три, некоторые авторы в базах являются суперпродуктивными, имея 241, 2015 и 192 публикации в базе eLibrary и 45, 29 и 28 публикаций в базе WoS.

[1] Согласно закону Лотки, число авторов, опубликовавших в течение определенного периода *n* статей, обратно пропорционально квадрату *n*. Этот закон можно проверить математической функцией.

Наиболее продуктивные авторы с наибольшим количеством работ по двум массивам приведены в табл. 6. Четверо из выделенных топ-20 авторов присутствуют в обеих базах, однако авторы из массива eLibrary, имеющие наибольшее количество публикаций, в базе WoS имеют 1,1 и 5 публикаций.

Таблица 6

АВТОРЫ С НАИБОЛЬШИМ КОЛИЧЕСТВОМ ПУБЛИКАЦИЙ, ПО ДВУМ МАССИВАМ WOS И ELIBRARY

Ранг	Массив WoS		Массив eLibrary	
	ID автора	кол-во публикаций	ID автора	кол-во публикаций
1	**TROTSUK_I**	45	429210_SI_Samygin_14461	241
2	PUZANOVA_Z	29	74486_SG_Maksimov_258_7082	215
3	KRAVCHEN_S	28	767943_TK_Rostovsk_924_1432_1488_4812_5350_13701	192
4	**TOSHCHEN_Z**	20	**137655_GE_Zborovsk_290_1255_7366_14141**	166
5	NARBUT_N	18	75266_NV_Dulina_306_1000	160
6	**ZBOROVSK_G**	18	145046_OE_Nojanzin_258_7082	150
7	SOROKIN_P	18	72232_JUG_Volkov_322_1432_3455_14829	147
8	SOKOLOV_M	18	129623_VA_Il'in_815	142
9	**GORSHKOV_M**	17	287431_AV_Verescha_322_1432_14461	133
10	YANITSKI_O	17	504328_MV_Morev_815	120
11	ROMANOVS_N	16	**251886_IV_Trotsuk_421_425**	120
12	TESLYA_A	15	495445_DA_Omel'che_258	119
13	KOZYREVA_P	14	**1382_ZHT_Toschenk_5_5350**	117

Окончание табл. 6

Ранг	Массив WoS		Массив eLibrary	
	ID автора	кол-во публикаций	ID автора	кол-во публикаций
14	SMIRNOV_A	14	73979_HV_Dzutsev_1432_4812	116
15	OBRAZTSO_I	14	442046_JUV_Stavropo_259_808	116
16	TIKHONOV_N	13	674856_NH_Gafiatul_322_761	111
17	GASPARIS_A	13	265785_VP_Babintse_340_1279_6227	110
18	RYBAKOVS_L	13	331427_SA_Il'inyh_1068	109
19	LAPIN_N	13	**72610_MK_Gorshkov_1432_14554**	109
20	LARINA_T	13	259120_PA_Ambarova_290	106

Примечание. Жирным шрифтом выделены фамилии авторов, встречающихся в топ-20 по обоим массивам; «ID автора» приведены согласно тому, как авторы указаны в соответствующем массиве данных.

Показатель исходящей центральности в сети **WA** показывает количество авторов в работах (табл. 7).

Таблица 7

КОЛИЧЕСТВО АВТОРОВ В ПУБЛИКАЦИЯХВ ДВУХ МАССИВАХ

WoS			eLibrary		
число авторов	*N*	доля от всех авторов, %	число авторов	*N*	доля от всех авторов, %
1	2217	62,29	1	49 973	66,43
2	844	23,71	2	18 473	24,55
3	327	9,19	3	5044	6,7
4	120	3,37	4	1144	1,52

Окончание табл. 7

WoS			eLibrary		
число авторов	*N*	доля от всех авторов, %	число авторов	*N*	доля от всех авторов, %
5	34	0,96	5	361	0,48
6	5	0,14	6	113	0,15
7	6	0,17	7	63	0,08
8	2	0,06	8	61	0,08
9	1	0,03			
12	1	0,03			
14	1	0,03			
15	1	0,03			

Максимальное число авторов в массиве WoS составляет 15; для массива eLibrary при сборе данных было установлено ограничение в 8 авторов. Как видно, доли публикаций статей с единственным автором в двух массивах являются практически идентичными – 62% в WoS и 66% в eLibrary. Это подтверждает обозначенную гипотезу о распространенности практики публикаций с единственным автором как части публикационной культуры в области социальных наук. Следующий самый часто встречающийся в публикациях формат – подготовка публикаций парами авторов – встречается в 24 и 25% статей в WoS и eLibrary соответственно; за ним следуют публикации, сделанные тремя (9% для WoS и 7% для eLibrary) и четырьмя (3% и 1.5%) авторами. Статьи с относительно большим количеством авторов встречаются в массивах в единичном виде.

Коллаборации авторов. На основе сети WA путем ее перемножения может быть построена базовая ненормализованная сеть соавторства Co, где сила связей рассчитывается исходя из количества публикаций, написанных авторами совместно, а петля обозначает общее количество работ у авторов, написанных в соавторстве и самостоятельно [4]. Доли авторов, не имеющих хотя бы одного

соавтора, для массивов eLibrary и WoS составляют 35,8 и 27,8% соответственно. Распределение по числу соавторов у авторов показывает, что большинство из них имеют одного (31% в eLibrary и 27% в WoS), двух (14% и 18,5%) или трех (6,5% и 12,5%) соавторов (рис. 6).

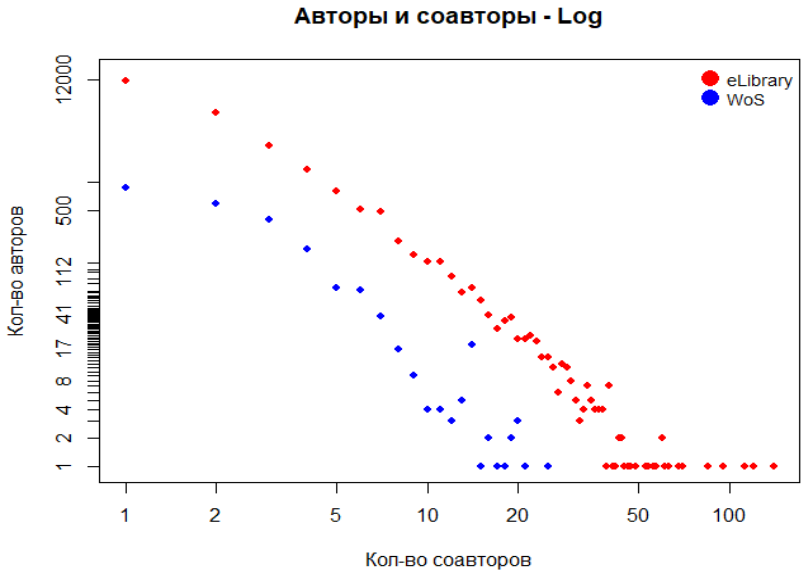


Рис. 6. Распределение количества соавторов по авторам в двух массивах данных (логарифмическая шкала)

Однако также выделяются авторы со значительным количеством соавторов, например: в массиве WoS – Н.Е. Покровский (25 соавторов), В.В. Щербина (21) и Ж.Т. Тощенко, Н.В. Романовский и А.Б. Гофман (20), в массиве eLibrary доля авторов с числом соавторов более 20 составляет 0,57%, или 212 авторов, среди которых лидирует С.И. Самыгин со 140 соавторами.

Работы и журналы. В табл. 8 приведены топ-25 журналов по количеству публикаций в двух массивах.

Таблица 8

ТОП-15 ЖУРНАЛОВ В ДВУХ МАССИВАХ ДАННЫХ ПО ЧИСЛУ ПУБЛИКАЦИЙ

Ранг	WoS			eLibrary		
	журнал	*N*	доля от всех публикаций, %	журнал	*N*	доля от всех публикаций, %
1	SOTSIOL ISSLED+	1923	54,0	Социологические исследования	1905	2,5
2	RUDN J SOCIOL	546	15,3	Экономика и социум	1759	2,3
3	SOCIOL OBOZR	309	8,7	Теория и практика общественного развития	1078	1,4
4	J ECON SOCIOL	245	6,9	Гуманитарные, социально-экономические и общественные науки	911	1,2
5	SOCIOL NAUK TEHNOL	167	4,7	Социально-гуманитарные знания	863	1,1
6	CHANG SOC PERSONAL	51	1,4	Социология	737	1,0
7	INT J SOCIOL SOC POL	37	1,0	Мониторинг общественного мнения: экономические и социальные перемены	731	1,0
8	COMP SOCIOL	23	0,6	Социология в современном мире: наука, образование, творчество	670	0,9

Окончание табл. 8

Ранг	WoS			eLibrary		
	журнал	*N*	доля от всех публикаций, %	журнал	*N*	доля от всех публикаций, %
9	INT J INTERCULT REL	22	0,6	Социальная политика и социология	630	0,8
10	SOC INDIC RES	20	0,6	Социальные и гуманитарные науки. Отечественная и зарубежная литература. Серия 11: Социология	598	0,8
11	FILOS-SOCIOL	18	0,5	Гуманитарий Юга России	582	0,8
12	SPORT SOC	8	0,2	Власть	528	0,7
13	POETICS	8	0,2	Журнал социологии и социальной антропологии	523	0,7
14	CORVINUS J SOCIOL PO	5	0,1	Общество: социология, психология, педагогика	494	0,7
15	CURR SOCIOL	5	0,1	Известия Саратовского университета. Новая серия. Серия: Социология. Политология	482	0,6

Лидером в обеих базах выступает журнал «Социологические исследования» – абсолютное количество публикаций в нем в WoS и в eLibrary примерно одинаково (1923 и 1905 соответственно). Если же посмотреть на вклад журнала в общее количество публикаций, то значение этого источника для базы WoS становится еще важнее – публикации в нем составляют 54% от всего массива данных. В eLibrary вклад «Социса» растворяется в связи с большим количеством журналов; несколько других журналов с большим вкладом идут с довольно небольшим отставанием. На основе распределения журналов из массива WoS видно, что вклад российских авторов в эту площадку (зарубежную «витрину») в основном делается через публикации в российских журналах, индексируемых в WoS (первые пять российских журналов составляют 90% публикаций). Однако при оценке вклада журналов нужно учитывать эффект периодичности (частоты выхода публикаций) и количества публикаций в каждом номере (которое обычно велико для недобросовестных журналов, которые могут присутствовать в базах).

Работы и ключевые слова. Показатель исходящей центральности в сети WK показывает количество ключевых слов в работе. Для работ из массива WoS этот показатель варьируется от 1 до 40, а из массива eLibrary – от 1 до 51 (при этом в 36,7% случаев значения пропущены). Показатель входящей центральности в сети WK показывает частоту использования различных ключевых слов в работах. Как показывает распределение этих значений для двух массивов (рис. 7), 77% ключевых слов в массиве eLibrary и 50,5% в WoS использованы только один раз, еще 10 и 14% соответственно – два раза, 3,6 и 7% – три раза и т.д.

В табл. 9 приведены топ-20 слов, наиболее часто используемых в обоих массивах. Повторяющиеся слова из двух массивов выделены цветом.

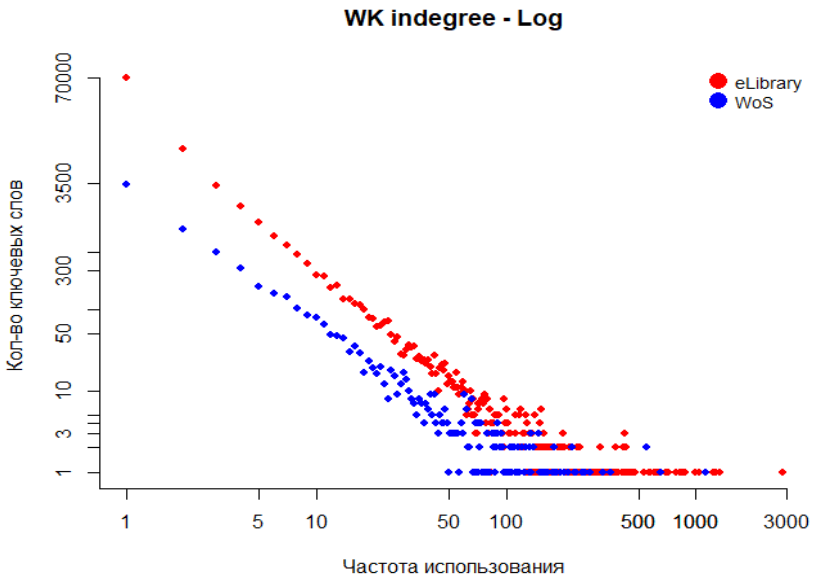


Рис. 7. Частота использования ключевых слов в работах в двух массивах данных (логарифмическая шкала)

Таблица 9

ТОП-20 КЛЮЧЕВЫХ СЛОВ ДЛЯ ОБОИХ МАССИВОВ

Ранг	eLibrary		WoS	
	слово	значение	слово	значение
1	youth	2849	social	1119
2	**society**	1343	russian	649
3	family	1269	**sociology**	547
4	values	1225	**russia**	547
5	**culture**	1035	**society**	353
6	**education**	990	sociological	322
7	globalization	876	analysis	279
8	students	855	theory	261
9	migration	839	study	253
10	socialization	820	**education**	233

Окончание табл. 9

Ранг	eLibrary		WoS	
	слово	значение	слово	значение
11	identity	791	**state**	232
12	civil society	716	political	231
13	modernization	696	research	230
14	communication	631	development	223
15	**state**	611	science	223
16	management	601	life	212
17	**russia**	580	**cultural**	202
18	region	579	value	193
19	internet	571	practice	185
20	**sociology**	534	public	182

Примечание. Полужирным шрифтом выделены слова, полностью повторяющиеся в двух списках, подчеркнуты слова, имеющие общую часть.

Выводы и обсуждение

Проведенный литературный обзор исследований по сравнению баз данных научных публикаций подтверждает, что даже при наличии альтернатив WoS является одним из самых популярных источников информации для наукометрических исследований. Безусловным плюсом работы с этой базой является наличие инструментов для выгрузки, предобработки и статистического и сетевого анализа публикаций, но, в случае с данными российских авторов, минусом – ограниченная представленность публикаций. В eLibrary, напротив, отечественные публикации представлены максимально полно (и не ограничиваются только публикациями в научных журналах и главами в монографиях); проблема заключается в отсутствии широко доступных сервисов по обработке и анализу данных для этой базы. В этой ситуации у исследователя, нацеленного на изучение современного состояния развития российской науки, возникает ряд вопросов: можно ли взять только одну базу в качестве источника информации, или необходимо комбинировать данные

из нескольких баз? в случае использования одной базы, насколько валидными будут полученные результаты? если данные должны комбинироваться, то как именно это нужно делать?

В нашем исследовании проводится сравнение двух баз через описательный анализ их возможностей по работе с данными и сопоставление двух массивов по одной и той же предметной области – социологии, – что является распространенной практикой в дизайне аналогичных исследований. Полученные массивы данных сравниваются по своей структуре, размеру, полноте метаданных, а также посредством анализа производных базовых сетей. Это важно не только с наукометрической точки зрения, но и с позиции изучения ориентаций ученых на международные и локальные научные сообщества, если думать о двух площадках как о двух возможных направлениях позиционирования ученых.

Несмотря на похожий набор метаданных, базы WoS и eLibrary имеют некоторые различия. Одним из преимуществ WoS является наличие списков литературы, что важно, если предполагается использовать анализ цитирований. В eLibrary авторы и организации имеют ID, однако по факту в большом количестве случаев эта информация отсутствует, что приводит к необходимости предварительной обработки массивов данных. В отличие от данных WoS, работа с которыми может осуществляться в нескольких программах, работа с данными eLibrary как по предобработке, так и по построению сетевых файлов для дальнейшего библиометрического анализа является гораздо более трудозатратой.

Рассмотренный массив eLibrary является гораздо более крупным, т.к. аккумулирует информацию из различных российских журналов и некоторых иностранных баз данных. WoS Core Collection имеет строгие критерии индексации журналов (что сокращает возможное число национальных журналов, которые могут быть представлены на этой площадке) и является только одной из баз, информация из которой должна включаться в eLibrary. Несмотря на то, что по логике формирования базы данных eLibrary рассматриваемые нами для примера массивы данных должны

в значительной степени пересекаться (массив WoS должен быть включен в массив eLibrary), пересечение между рассматриваемыми массивами является далеко не полным (около 30% массива WoS входят в массив eLibrary). Это может объясняться как несовершенством реализованной процедуры поиска идентичных публикаций и авторов, так и тем, что в массивах содержатся уникальные публикации и авторы. Эта часть анализа в настоящий момент может рассматриваться как экспериментальная и заслуживает дальнейшей проработки для уточнения пересечения массивов. Отметим, что наличие DOI у всех статей и ResearcherID/ORCID-ID у авторов могло бы существенно упростить эту задачу.

Динамика количества публикаций в двух массивах показывает, что база WoS прирастает более активно. Однако данные в обоих массивах распределяются похожим образом, что говорит о том, что они следуют похожим библиометрическим трендам и законам. Схожим образом в обоих массивах разделяются доли числа работ у авторов (две трети авторов с одной статей), авторов у работ (две трети работ наблюдаемых в социальных науках «авторов-одиночек», четверть работ, написанных в парах), соавторов у авторов (около трети авторов без соавторов и столько же – с одним соавтором); доля статей с одним ключевым словом в массиве eLibrary выше, чем в WoS (77% против 50%). Выделенные топ-единицы анализа при этом пересекаются только частично, что говорит о наличии своих особенностей в каждом массиве – самых продуктивных авторов для каждой площадки, наиболее используемых журналов и уникальных ключевых слов, характеризующих исследования. Более подробный анализ наблюдаемых пересечений и отличий может помочь ответить на различные содержательные вопросы о специфике исследований, ориентированных на разные аудитории (хотя возникает вопрос, насколько «ориентированными» на зарубежные исследовательские группы являются публикации в WoS, изначально вышедшие в российских журналах и внесенные в базу благодаря их индексации).

На основе проделанного сравнения и изучения логики формирования рассмотренных баз данных научных публикаций можно сделать некоторые выводы и рекомендации по поводу выбора базы данных для анализа публикаций российских авторов. Выяснилось, что множество статей в eLibrary не идентично множеству статей РИНЦ – в последнем индексируется меньше журналов и иных типов публикаций ввиду применения более строгих правил отбора, что делает данные в базе РИНЦ более надежными для анализа развития научной продукции. Обращение в качестве источника данных к базе RSCI, применяющей строгие критерии для отбора топовых российских журналов, сужает объем изучаемого числа публикаций, предоставляя доступ только к статьям, опубликованным в российских журналах. Обращение к WoS CC хотя и дает доступ к публикациям российских авторов в зарубежных журналах, включает публикации лишь из некоторых российских журналов, поэтому ограничивает анализируемый объем статей в наибольшей степени. Конечный выбор той или иной базы должен диктоваться исследовательской задачей: изучение представленности российских авторов в международном пространстве, очевидно, требует обращения к базе WoS (однако стоит рассматривать и выход за пределы ее ядра CC), а анализ сугубо российских публикаций может быть осуществлен путем анализа базы RSCI. Вместе с тем для изучения трендов развития российской науки в целом и практики ее воспроизводства стоит обратиться к базе РИНЦ или eLibrary. В идеальной ситуации анализ публикаций российских авторов должен осуществляться на основе нескольких баз данных, однако методологические вопросы выгрузки данных, поиска совпадений между базами и их объединения в единый массив пока являются открытыми и требуют дальнейшей разработки.

В качестве общей рекомендации нужно сказать, что исследователь, работающий в области библиометрического анализа, должен хорошо понимать структуру баз, с которыми он работает, чтобы получить нужную ему информацию на входе для дальней-

шего анализа, а не просто «искать где светлее» (например, в WoS, так как разработаны инструменты для анализа), а также ставить исследовательские вопросы с пониманием ограничений в покрытии баз данных. С точки зрения получения валидных результатов важной является также оценка полноты отдельных метаданных в библиографических описаниях.

Безусловно, важным вопросом является доступность баз данных. К сожалению, рассмотренная в рамках проведенного анализа база WoS с недавних пор недоступна российским исследователям, а доступ к API-сервису eLibrary отсутствует у большинства исследователей. На наш взгляд, наличие функциональных возможностей для сбора данных или открытого доступа к API-сервису eLibrary может стать важным условием для развития библиометрического анализа публикаций российских авторов и наукометрических исследований в российской практике в целом. Еще одним вариантом для получения данных может стать использование альтернативных открытых баз данных научных публикаций (например, базы OpenAlex, выбор которой уже приходит на замену традиционным базам WoS и Scopus в некоторых научно-образовательных организациях).

Дальнейшая работа над этой тематикой требует разработок в области методологии сбора, поиска совпадений и объединения массивов из различных источников. Полученные результаты в области социологии интересно сравнить с другими предметными областями. Помимо наукометрического интереса, анализ и сравнение публикационной активности исследователей на разных – международных и отечественных – площадках может помочь ответить и на многие содержательные вопросы, возникающие при изучении локальных научных сообществ, которые находятся за пределами рассмотрения данной статьи.

ЛИТЕРАТУРА

1. *Bar-Ilan J.* Informetrics at the beginning of the 21st century – A review // Journal of informetrics. 2008. Vol. 2. P. 1–52. DOI: 10.1016/j.joi.2007.11.001. EDN: MISIBR.

2. *Mingers J., Leydesdorff L.* A review of theory and practice in scientometrics // European journal of operational research. 2015. Vol. 246, № 1. P. 1–19. DOI: 10.1016/j.ejor.2015.04.002. EDN: UQPVRP.

3. *Rousseau R., Egghe L., Guns R.* Becoming metric-wise: A bibliometric guide for researchers / Ed. by W. Glänzel [et al.]. Cambridge, MA: Chandos Publishing, 2018. 850 p. ISBN: 0081024754, 9780081024751.

4. Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution / V. Batagelj, P. Doreian, A. Ferligoj, N. Kejžar. Hoboken, NJ: WileyBlackwell, 2014. 464 p. ISBN: 978-1-118-91537-0.

5. *Моисеев С.П., Мальцева Д.В.* Отбор источников для систематического обзора литературы: сравнение экспертного и алгоритмического подходов // Социология: методология, методы, математическое моделирование (Социология: 4М). 2018. № 47. С. 7–43. EDN: MZXVXW.

6. *Булычева Е.Е., Мальцева Д.В.* Выделение актуальных тематик в социологии: взгляд сквозь призму анализа сети цитирований // Мониторинг общественного мнения: экономические и социальные перемены. 2020. № 6 (160). С. 113–140. DOI: 10.14515/monitoring.2020.6.971. EDN: UGIDGS.

7. *Harzing A.W., Alakangas S.* Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison // Scientometrics. 2016. Vol. 106. P. 787–804. DOI: 10.1007/s11192-015-1798-9. EDN: ZGNBBS.

8. The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis / V.K. Singh, P. Singh, M. Karmakar [et al.] // Scientometrics. 2021. Vol. 126. P. 5113–5142. DOI: 10.1007/s11192-021-03948-5. EDN: FLHAPG.

9. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations / A. Martín-Martín, M. Thelwall, E. Orduna-Malea, E. Delgado López-Cózar // Scientometrics. 2021. Vol. 126, № 1. P. 871–906. DOI: 10.1007/s11192-020-03690-4. EDN: XNWTDQ.

10. *Harzing A.W*. Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? // Scientometrics. 2019. Vol. 120, № 1. P. 341–349. DOI: 10.1007/s11192-019-03114-y. EDN: VKFCPM.

11. *Zhu J., Liu W.* A tale of two databases: The use of Web of Science and Scopus in academic papers // Scientometrics. 2020. Vol. 123, № 1. P. 321–335. DOI: 10.1007/s11192-020-03387-8. EDN: LZMVNM.

12. *Gusenbauer M.* Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases // Scientometrics. 2019. Vol. 118, № 1. P. 177–214. DOI: 10.1007/s11192-018-2958-5. EDN: ECWMGT.

13. *Moed H.F., Markusova V., Akoev M.* Trends in Russian research output indexed in Scopus and Web of Science // Scientometrics. 2018. Vol. 116. P. 1153–1180. DOI: 10.1007/s11192-018-2769-8. EDN: VBDLNY.

14. *Vera-Baceta M.A., Thelwall M., Kousha K.* Web of Science and Scopus language coverage // Scientometrics. 2019. Vol. 121, № 3. P. 1803–1813. DOI: 10.1007/s11192-019-03264-z. EDN: IHLJRA.

15. *Ruiz-Pérez R., López-Cózar E.D., Jiménez-Contreras E.* Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies // Journal of the medical library association. 2002. Vol. 90, № 4. P. 411–430.

16. *Adriaanse L.S., Rensleigh C.* Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison // The Electronic Library. 2013. Vol. 31, № 6. P. 727–744. DOI: 10.1108/EL-12-2011-0174.

17. *Еременко Г.О.* Сравнение уровня публикаций российских ученых в базах данных Web of Science, Scopus и RSCI: статья в открытом архиве // НЭБ. 28.02.2020. URL: https://elibrary.ru/wos_scopus_rsci.asp (дата обращения: 01.12.2023). EDN: CQMPRA.

18. Russian index of science citation: Overview and review / O. Moskaleva, V. Pislyakov, I. Sterligov [et al.] // Scientometrics. 2018. Vol. 116. P. 449–462. DOI: 10.1007/s11192-018-2758-y. EDN: XTIRDN.

19. The Russian Science Citation Index (RSCI): the first three years (2016–2018) / S. V. Gorin, A. M. Koroleva, A. N. Gerasimov, A. A. Voronov // European Science Editing. 2020. Vol. 46. DOI: 10.3897/ese.2020.e51051. EDN: XDXXDQ.

20. *Мальцева Д.В., Ващенко В.А., Капустина Л.В.* Методология обработки библиографических данных на русском языке для построения сетей коллаборации (на примере базы данных eLibrary) // Социология: методология, методы, математическое моделирование (Социология: 4М). 2022. № 54–55. С. 45–78. DOI: https://doi.org/10.19181/4m.2022.31.1-2.2. EDN: GRRLBQ.

21. *Batagelj V.* WoS2Pajek. Networks from web of science. Version 1.5 (2017). URL: http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:wos2pajek (дата обращения: 01.12.2023).

Приложение

АНАЛИЗ БАЗ ДАННЫХ, ИНДЕКСИРОВАННЫХ В ELIBRARY

В массиве eLibrary для каждой публикации содержится информация о том, в каких еще наукометрических базах (РИНЦ, RSCI, WoS, Scopus, ВАК) она проиндексирована. Это дает возможность посмотреть на распределение и пересечение публикаций в разных базах на основе информации, предоставляемой площадкой eLibrary.

Как видно из табл. 1, большинство публикаций из массива eLibrary представлены в журналах, индексируемых в базе РИНЦ (85%) и входящих в список ВАК (53%). Значительно меньшее количество статей опубликованы в журналах, индексируемых в RSCI (11%), Scopus (8%) и WoS CC (7%). Безусловно, множества, составляемые массивами публикаций в разных базах, являются пересекающимися (статьи могут входить в разные базы). На основе имеющихся данных о принадлежности статей в массиве eLibrary к разным базам можно посмотреть на пересечение (как множество общих единиц) и объединение (как множество всех единиц) между различными базами на уровне публикаций и журналов.

Таблица 1

ИНДЕКСАЦИЯ ПУБЛИКАЦИЙ ИЗ МАССИВА ELIBRARY В РАЗЛИЧНЫХ БАЗАХ

База	Показатель	Входит в базу	Не входит в базу	Сумма
РИНЦ/RISC	абсолютные значения	65 103	10 129	75 232
	доля, %	87	13	100
ВАК	абсолютные значения	40 046	35 186	75 232
	доля, %	53	47	100
RSCI	абсолютные значения	8 179	67 053	75 232
	доля, %	11	89	100

Окончание табл. 1

База	Показатель	Входит в базу	Не входит в базу	Сумма
Scopus	абсолютные значения	6 222	69 010	75 232
	доля, %	8	92	100
WoS	абсолютные значения	5 226	70 006	75 232
	доля, %	7	93	100

Сходство баз на уровне публикаций. Разные комбинации баз данных составляют разные доли от общего числа статей в массиве eLibrary (табл. 2).

● Некоторые включенности одних множеств в другие объясняются известной информацией о создании баз: так, все публикации из RSCI по определению включены в базу РИНЦ, поскольку являются подмножеством статей, опубликованных в российских топ-журналах.

● Известно, что база РИНЦ включает публикации российских авторов, представленные в журналах WoS и Scopus, публикации в массиве, входящие в эти базы, также полностью (5226 для WoS) и почти полностью (6212 из 6222 для Scopus) входят в РИНЦ.

● Ситуация для базы RSCI несколько иная: из 8179 публикаций в этой базе российских топ-журналов в журналах WoS CC также индексированы 4263 работ (52%), а из 5226 публикаций в WoS 963 статьи (18,4%) не входят в RSCI, но индексируются только в WoS CC (и попадают в базу благодаря тому, что их индексирует РИНЦ).

● Число статей из базы RSCI, также индексированных в базе Scopus, составляет 5249 (64,2% от всех публикаций в RSCI), а число уникальных статей из Scopus в нашей базе составляет 973 статьи (15,6% от всех публикаций в Scopus).

● Общее пересечение статей из собранного массива данных (75 232 публикации), входящих, по данным eLibrary, и в WoS, и в Scopus, составляет 4170 статей – что составляет 79,8% от всех публикаций WoS в массиве и 67% от всех публикаций в Scopus.

● Аналогичная доля рассчитывается и на пересечении этих двух баз и РИНЦ (опять же, по природе создания базы), но если сравнивать с базой RSCI, то число общих статей на пересечении трех баз составляет 3875 (что составляет 74,1% от всех статей в WoS, 62,3% – в Scopus и 47,4% – в RSCI).

● Общее число публикаций, представленных во всех трех базах (RSCI, Scopus, WoS), которые составляют ядро РИНЦ, рассчитанное как объединение множеств, составляет 9820 публикаций – 13% от всех публикаций в собранном массиве (75 232 публикации).

Обращает на себя внимание интересный факт: табл. 1 показывает, что не все статьи, вошедшие в массив eLibrary, входят в базу РИНЦ – только 87%. Предполагая, что оставшиеся 13% статей распределены по другим базам, указанным в массиве eLibrary, мы посмотрели на объединения баз Scopus, WoS и РИНЦ, а также объединение всех пяти баз (табл. 2). Выяснилось, что первое объединение составляет 65 113 публикаций, а второе – 65 308 публикаций – то есть снова около 87% публикаций из базы. Оставшиеся 9924 статей не входят ни в одну из пяти баз, указанных в eLibrary.

Таблица 2

СХОДСТВО МЕЖДУ БАЗАМИ ДАННЫХ ПО ЧИСЛУ ПУБЛИКАЦИЙ (МАССИВ ELIBRARY)

База	Число публикаций	Доля от общего числа статей, %
Пересечение (множество общих статей – правило «И»)		
РИНЦ + RSCI	8179	10,9
РИНЦ + WoS	5226	6,9
РИНЦ + Scopus	6212	8,3
RSCI + WoS	4263	5,7
RSCI + Scopus	5249	6,98
WoS + Scopus	4170	5,5
Scopus + WoS + РИНЦ	4170	5,5
Scopus + WoS + RSCI	3875	5,2
РИНЦ + ВАК	39 841	52,6

Окончание табл. 2

База	Число публикаций	Доля от общего числа статей, %
Объединение (множество всех статей – правило «ИЛИ»)		
Scopus + WOS + RSCI (ядро РИНЦ)	9820	13
Scopus + WOS + РИНЦ	65 113	86,5
Scopus + WOS + RSCI + ВАК	40 292	53,6
Scopus + WOS + RSCI + ВАК + РИНЦ	65 308	86,8
РИНЦ + ВАК	65 308	86,8

Более внимательный анализ этого подмассива работ показал, что они опубликованы в журналах, с которыми заключено лицензионное соглашение на размещение издания на eLibrary.ru. Кроме того, выборочный анализ некоторых журналов с помощью системы SCIENCE INDEX на eLibrary показал, что в определенные периоды времени эти журналы индексировались в РИНЦ. Топ журналов из данного подмассива приведен в табл. 3; ведущим источником выступает «Экономика и социум» с 1759 статьями (срочные платные публикации). Таким образом, в ходе анализа была уточнена реализованная стратегия сбора данных, осуществленная ООО «НЭБ»: отбор статей, индексируемых eLibrary, не аналогичен отбору по статьям, индексируемым в РИНЦ.

Таблица 3

ЖУРНАЛЫ С НАИБОЛЬШИМ КОЛИЧЕСТВОМ СТАТЕЙ ИЗ ПОДМАССИВА ПУБЛИКАЦИЙ, ИНДЕКСИРУЕМЫХ ТОЛЬКО В ELIBRARY

№	Название журнала	Кол-во статей	№	Название журнала	Кол-во статей
1	Экономика и социум	1759	12	Студенческий	166
2	Молодой ученый	441	13	Гуманитарные научные исследования	145

Окончание табл. 3

№	Название журнала	Кол-во статей	№	Название журнала	Кол-во статей
3	Сборники конференций НИЦ Социосфера	425	14	Современные тенденции развития науки и технологий	137
4	Аллея науки	327	15	Студенческий вестник	122
5	NovaInfo.Ru	250	16	Научный альманах	119
6	Вестник современных исследований	193	17	Вестник научных конференций	116
7	Теория и практика современной науки	176	18	Евразийский союз ученых	115
8	Актуальные проблемы гуманитарных и естественных наук	176	19	Современные научные исследования и инновации	110
9	Стратегия устойчивого развития регионов России	174	20	Форум молодых ученых	106
10	Система ценностей современного общества	171	21	Colloquium-journal	106
11	Сборник научных трудов SWorld	168	22	Альманах современной науки и образования	99

Еще одно пересечение баз данных относится к публикациям, входящим в список ВАК. По данным eLibrary, всего в собранном массиве из 75 232 работ 40 046 публикаций входят в эту базу и их подавляющее большинство (99,5%) входит в РИНЦ; разница между базами составляет 205 журналов. При объединении же

множества ВАК со Scopus, WoS и RSCI получается 40 292 публикации – всего на 246 больше, чем в базе ВАК. Получается, что список ВАК в значительной степени состоит из журналов, индексируемых в этих трех базах (ядре РИНЦ). Результат объединения пяти баз в числовом выражении аналогичен объединению множеств РИНЦ и ВАК – то есть все статьи, индексируемые в ядре РИНЦ, входят в эти два множества.

Сходство баз на уровне журналов. Количество журналов, индексируемых в разных базах по массиву eLibrary (на основании представленных площадкой данных), показано в табл. 4.

Таблица 4

ИНДЕКСАЦИЯ ЖУРНАЛОВ ИЗ МАССИВА ELIBRARY В РАЗЛИЧНЫХ БАЗАХ

База	Количество журналов	Доля от общего числа журналов, %
РИНЦ	3580	91,56
RSCI	202	5,17
Scopus	193	4,94
WoS	148	3,79
ВАК	1310	33,5
Всего журналов	3910	100

Подавляющее число источников (91,6%), в которых опубликованы статьи в базе, индексируются в РИНЦ; доли журналов, индексируемых в базах RSCI, Scopus, WoS, небольшие и составляют 4–5% от всех журналов. Треть всех журналов включены в список ВАК.

По анализируемому массиву данных из всех журналов в РИНЦ, где опубликованы работы по социологии, в число топ-журналов, отобранных для базы RSCI, входит 202 журнала. Число журналов из РИНЦ, индексируемых в WoS, составляет 148 журналов, а в Scopus – 193 журнала; на пересечении эти две зарубежные

базы дают в РИНЦ 92 журнала (что составляет 48% от всех журналов в Scopus и 62% от всех журналов в WoS). Число журналов из RSCI на пересечении с WoS дает 51 журнал, а со Scopus – 83, на пересечении трех баз находится 41 журнал.

Полученные результаты демонстрируют, что база РИНЦ, максимально близкая по размеру базе eLibrary (но не полностью покрывающая ее), в значительной степени пересекается с другими, меньшими по размеру базами библиографических данных, в том числе с базой WoS. В тексте статьи поиск такого пересечения производится через сравнениие двух рассматриваемых массивов.

Maltseva Daria V.,
Candidate of Sciences in Sociology, Head of the International Laboratory for Applied Network Research, HSE University, Moscow, Russia, dmaltseva@hse.ru

Pavlova Irina A.,
Candidate of Sciences in Economics, Deputy Head of the International Laboratory for Applied Network Research, HSE University, Moscow, Russia, iapavlova@hse.ru

Kapustina Lika V.,
Research Assistant at the International Laboratory for Applied Network Research, HSE University, Moscow, Russia, lkapustina@hse.ru

Vashchenko Vasilisa A.,
R*esearch Assistant at the International Laboratory for Applied Network Research, HSE University, Moscow, Russia,* vvashchenko@hse.ru

Fiala Dalibor,
Associate Professor at the Faculty of Applied Sciences, Department of Computer Science and Engineering, West Bohemian University, Czech Republic, Pilsen, dalfia@kiv.zcu.cz

Comparative analysis of the capabilities of WoS and eLibrary for analyzing bibliographic networks

This article presents a comparative analysis of two major scientific publication databases: Web of Science Core Collection and eLibrary – to identify their differences and unique opportunities for exploration of bibliographic networks of Russian scientific authors. Current shortage of tools and approaches for collection, processing and analysis of bibliographic data in the Russian language constitutes the relevance of this study. Empirical analysis is based on comparison of respective arrays of scientific publications in the field of sociology over the period of 2010-2021. We propose a set of comparison criteria including those related to the procedure of data access, quality of data management, quantitative and qualitative features of the data. Inspection of the databases based on the proposed criteria aids in identification of intersections between both the collections and the respective qualitative observations about them. We make conclusions regarding the comparative advantages and weaknesses of both databases in regards to their potential as the sole data source for bibliographic studies, and make recommendations for their

effective use in research on Russian science.
Keywords: network analysis, comparative analysis, bibliographic databases, bibliographic networks, eLibrary, Web of Science

References

1. Bar-Ilan J. Informetrics at the beginning of the 21st century – A review, *Journal of informetrics*. 2008, vol. 2, p. 1–52. DOI: 10.1016/j.joi.2007.11.001.
2. Mingers J., Leydesdorff L. A review of theory and practice in scientometrics, *European journal of operational research*, 2015, vol. 246, no. 1, p. 1–19. DOI: 10.1016/j.ejor.2015.04.002.
3. Rousseau R., Egghe L., Guns R. *Becoming metric-wise: A bibliometric guide for researchers*, ed. by W. Glänzel [et al.]. Cambridge, MA: Chandos Publishing, 2018. 850 p. ISBN: 0081024754, 9780081024751.
4. Batagelj V., Doreian P., Ferligoj A., Kejžar N. *Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution*. Hoboken, NJ: WileyBlackwell, 2014, 464 p. ISBN: 978-1-118-91537-0.
5. Moiseev S.P., Maltseva D.V. Source selection for systematic literature reviews: a comparison of expert and algorithmic approaches (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2018, no. 47, p. 7–43.
6. Bylucheva E.E., Maltseva D.V. Identifying relevant topics in sociology: a bibliographic network analysis view (in Russian), *Monitoring of Public Opinion: Economic and Social Changes*, 2020, no. 6 (160), p. 113–140. DOI: 10.14515/monitoring.2020.6.971.
7. Harzing A.W., Alakangas S. Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison, *Scientometrics*, 2016, vol. 106, p. 787–804. DOI: 10.1007/s11192-015-1798-9.
8. Singh V.K., Singh P., Karmakar M. [et al.] The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis, *Scientometrics*, 2021, vol. 126, p. 5113–5142. DOI: 10.1007/s11192-021-03948-5.
9. Martín-Martín A., Thelwall M., Orduna-Malea E., Delgado López-Cózar E. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations, *Scientometrics*, 2021, vol. 126, no. 1, p. 871–906. DOI: 10.1007/s11192-020-03690-4.

10. Harzing A.W. Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics*, 2019, vol. 120, no. 1, p. 341–349. DOI: 10.1007/s11192-019-03114-y.
11. Zhu J., Liu W. A tale of two databases: The use of Web of Science and Scopus in academic papers, *Scientometrics*, 2020, vol. 123, no. 1, p. 321–335. DOI: 10.1007/s11192-020-03387-8.
12. Gusenbauer M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases, *Scientometrics*, 2019, vol. 118, no.1, p. 177–214. DOI: 10.1007/s11192-018-2958-5.
13. Moed H.F., Markusova V., Akoev M. Trends in Russian research output indexed in Scopus and Web of Science, *Scientometrics*, 2018, vol. 116, p. 1153–1180. DOI: 10.1007/s11192-018-2769-8.
14. Vera-Baceta M.A., Thelwall M., Kousha K. Web of Science and Scopus language coverage, *Scientometrics*, 2019, vol. 121, no. 3, p. 1803–1813. DOI: 10.1007/s11192-019-03264-z.
15. Ruiz-Pérez R., López-Cózar E.D., Jiménez-Contreras E. Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies, *Journal of the medical library association*, 2002, vol. 90, no. 4, p. 411–430.
16. Adriaanse L.S., Rensleigh C. Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison, *The Electronic Library*, 2013, vol. 31, no. 6, p. 727–744. DOI: 10.1108/EL-12-2011-0174.
17. Eremenko G.O. Comparing publication levels of Russian scientists across Web of Science, Scopus and RSCI databases (in Russian), *NAB (Scientific Electoring Library)*, 28.02.2020, URL: https://elibrary.ru/wos_scopus_rsci.asp (date of access: 01.12.2023).
18. Moskaleva O., Pislyakov V., Sterligov I. [et al.] Russian index of science citation: Overview and review, *Scientometrics*, 2018, vol. 116, p. 449–462. DOI: 10.1007/s11192-018-2758-y.
19. Gorin S.V., Koroleva A.M., Gerasimov A.N., Voronov A.A. The Russian Science Citation Index (RSCI): the first three years (2016–2018), *European Science Editing*, 2020, vol. 46. DOI: 10.3897/ese.2020.e51051.
20. Maltseva D.V., Vashchenko V.A., Kapustina L.V. Methodology of processing bibliographic data in Russian language to construct collaboration networks (using the example of the eLibrary database) (in Russian),

Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling), 2022, no. 54–55, p. 45–78. DOI: 10.19181/4m.2022.31.1-2.2.

21. Batagelj V. *WoS2Pajek. Networks from web of science*, Version 1.5 (2017). URL: http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:wos2pajek (date of access: 01.12.2023).