

Международный научно-исследовательский отчет «Ценность наукометрических баз данных: от поиска информации — к глубокой аналитике»

Джонатан Адамс (Jonathan Adams), Дэвид Пендлбери (David Pendlebury) и Мартин Шомшор (Martin Szomszor)

Биографии авторов

Джонатан Адамс занимает должность главного научного сотрудника Института научной информации (ISI). Кроме того, он является штатным профессором в Институте политики Королевского колледжа Лондона. В 2017 году Эксетерский университет удостоил его звания почетного доктора наук за его работу в области высшего образования и научной политики.

Дэвид Пендлбери является руководителем отдела научного анализа в ISI. С 1983 года он использует данные *Web of Science* для изучения структуры и динамики исследований. В течение многих лет он работал с основателем ISI Юджином Гарфилдом (Eugene Garfield). Совместно с Генри Смоллом (Henry Small) Дэвид разработал *Essential Science Indicators* ISI.

Мартин Шомшор занимает должность директора Института научной информации. Кроме того, он был руководителем отдела научной аналитики в ISI. В 2015 году он был назван одним из 50 лучших специалистов по работе с информацией в рамках премии «Век информации» в Великобритании за свою работу по созданию базы данных научного влияния REF2014 для Совета по финансированию высшего образования Англии (HEFCE).

Фундаментальное прошлое, многообещающее будущее

Институт научной информации

ISI опирается на работу доктора Юджина Гарфилда — своего первоначального основателя и пионера информетрии. ISI был восстановлен в 2018 году и назван в честь компании, которая предшествовала *Web of Science Group* и которую он основал. В институте проводятся аналитические экспертизы, которые основаны на его наследии и используют достижения научно-технического прогресса.

Наша международная команда признанных в своих областях экспертов сосредоточена на развитии существующих и разработке новых наукометрических и аналитических подходов, а также на продвижении сотрудничества с партнерами и коллегами-исследователями в мировом научном сообществе.

Сейчас, будучи «университетом» *Web of Science Group*, ISI:

- поддерживает базовые знания и редакционную строгость, на которых строится указатель научного цитирования *Web of Science* и связанные с ним продукты и услуги. В течение почти полувека наши оценка и отбор информации основываются на научных исследованиях и объективном анализе. Благодаря тщательно отобранному, структурированному и полным данным *Web of Science* даёт возможность получить ясное представление о научном вкладе и ценности самых влиятельных научных журналов в мире. Эта экспертная оценка позволяет исследователям, издателям, редакторам, библиотекарям и грантодателям исследовать ключевые факторы ценности журнала для разных аудиторий, наилучшим образом используя широкий массив доступных данных и метрик.

- проводит исследования для поддержания, расширения и улучшения базы знаний и с помощью наших докладов и публикаций, а также мероприятий и конференций распространяет эти знания среди наших коллег, партнеров и всех тех, кто занимается научными исследованиями в университетах, коммерческих компаниях, финансирующих организациях, издательствах и государственных структурах.

Информационная справка

В 2019 году около 145 000 исследователей из 139 стран, работающих в различных исследовательских областях, ежедневно проводили в *Web of Science* поиск информации для изучения и поиска научных источников и литературы, имеющих отношение к их текущему исследованию.

В 1981 году в *Web of Science* были проиндексированы около 500 000 публикаций (значимых научных статей и обзоров) из 6800 журналов. В 2019 году их число значительно увеличилось до 2,5 млн публикаций из 21 300 журналов. Это обширный источник данных для широкого спектра аналитических задач.

Однако существует мало исследований на тему того, как ещё можно использовать *Web of Science* в качестве библиографической базы данных не только для поиска и нахождения информации. Наш анализ показывает, что *Web of Science* по сути является основным источником данных о публикациях и цитировании для большинства систематических обзоров исследований по широкому кругу дисциплин, а также содержит примерно вдвое больше исследований на тему организации и оценки научных исследований по сравнению с любым другим источником. *Web of Science* является основным источником данных для подобной работы в США, Китае и большинстве стран Западной Европы. А учёные из Ирана, Италии и Австралии предпочитают в качестве такого источника базу данных Scopus.

Основными бенефициарами структурированного использования библиографических записей *Web of Science* являются биомедицинские исследователи, которые привыкли на регулярной основе использовать исходные данные исследований для обзоров разработки и текущего состояния тем исследований, имеющих решающее значение

для здоровья человека и борьбы с болезнями. Актуальная структура таких публикаций демонстрирует критическую значимость библиографических материалов *Web of Science* для обзора политики в области здравоохранения, направленной на решение таких проблем, как заболеваемость раком, охрана здоровья женщин и сердечно-сосудистые заболевания, а также для управления медицинскими результатами и разработки инновационных подходов и методов лечения.

Отслеживание таких публикаций во времени помогает выявлять новые области и определять направление финансирования исследований. Если мы понимаем, кто наиболее активно публикуется в определённой области, то мы можем также увидеть, как распределяется экспертное знание по странам и организациям, что, в свою очередь, помогает сравнить и оценить научные достижения и политику этих стран и организаций.

Наш анализ показывает, что *Web of Science* действительно является основным источником публикаций и цитирования для большей части системной научной аналитики.

Введение

В 2019 году поисковые запросы в *Web of Science* ежедневно производили около 145 000 исследователей из 139 стран, представляющих весь спектр научных дисциплин в области естественных, социальных и, все больше, гуманитарных наук.

Учёные используют *Web of Science* для планирования новых исследований и формулирования вопросов, на которые им необходимо найти ответ, а также для поиска и нахождения важнейших библиографических материалов, необходимых в рамках текущих исследований.

Удивительно, но почти никто не исследует, как научное сообщество использует эту библиографическую базу данных помимо целей поиска информации. В 2008 г Прингл (Pringle) исследовал всё более активное использование библиографических метаданных при оценке исследований, а в 2018 г Шнель (Schnell) отметил историческую роль *Web of Science* как первой базы данных научного цитирования для анализа данных и наукометрии. Совсем недавно Роллинз и Ян Ли (Rollins Li/Yan Li, 2018 г) подтвердили, что количественное влияние *Web of Science* не было тщательно изучено в ходе научных исследований. Они исследовали способы упоминания этого источника информации в выборке из 19 478 статей, опубликованных в период с 1997 по 2017 гг, и проанализировали распределение по странам, учреждениям и областям исследований. По всей видимости, это было первое исследование, в рамках которого проводилось эмпирическое исследование использования базы данных в таких документах.

В этом отчете мы развили идеи из исследования Li et al (2018) и Schnell (2018), сосредоточив внимание на методах использования данных *Web of Science* исследователями для обнародования результатов своих исследований по всей системе научно-исследовательской коммуникации, а также на способах использования этой информации для поддержки конкретных исследований, особенно систематических обзоров, и для улучшения системы управления исследованиями. За последние

несколько десятилетий число статей в *Web of Science* и охват исследований значительно расширились. В 1981 году в ней ежегодно индексировалось около 500 000 научных статей (большие научные статьи и обзоры) из приблизительно 6800 журналов. В настоящее время индексируется около 2,5 млн статей из приблизительно 21 300 журналов. В базе данных содержится около 20 миллионов статей, написанных исследователями из США или около 27 миллионов статей, написанных учёными из Европейского Союза, что делает её огромным массивом данных для широкого спектра аналитических задач.

За последние несколько десятилетий число статей в *Web of Science* и охват исследований значительно расширились.

Записи публикаций структурированы по категориям на основе журналов (изначально предназначенных для облегчения поиска), но теперь чаще используются в качестве основы для сравнительной аналитики. Существует 254 детализированные категории журналов *Web of Science* (по состоянию на январь 2020 года) и 21 более общая категория в инструменте Essential Science Indicators. Анализ цитирования часто проводится с использованием категорий *Web of Science*, и тесная дисциплинарная связь между журналами, которые относятся к той или иной категории, означает, что среднее количество цитирований для всех статей в этой категории в конкретном году является значимым международным критерием научного качества.

Однако управление научными исследованиями редко происходит на этом уровне категорий. Гораздо вероятнее, что оно будет производиться на более детальном уровне, внутри или между установленными категориями журналов. Исследования становятся все более междисциплинарными, поскольку тематика исследований часто охватывает традиционные дисциплинарные структуры и самые современные решения основных проблем, затрагивающих множество областей. Это означает, что информация, необходимая как для исследователей, так и для тех, кто занимается отбором, финансированием и оценкой результатов исследовательских проектов, вероятно, будет получена в результате выполнения целенаправленных поисков, анализов и отчетов, и будет представлять широкий интерес для всей области.

В массиве данных, собранных в базе, можно найти ответы на множество вопросов. Объединение в единой структуре связанных публикаций даёт исходные данные для обзоров, которые раскрывают развитие и текущее состояние более широких тем исследований. Отслеживание библиографических материалов во времени выявить новые области и может помочь в определении наиболее эффективных с точки зрения финансирования направлений. Определение наиболее активных авторов в избранной области показывает распределение экспертных знаний и поддерживает сравнительную оценку деятельности и результатов.

Когда, что и где?

Первое, что мы делаем в этом отчете, это определение количества статей (т.е. как статей, так и обзоров), публикуемых каждый год и индексируемых в *Web of Science*, авторы которых также ссылаются в заголовке, аннотации и ключевых словах на анализ данных из *Web of Science* или других широко используемых международных баз данных научного цитирования.

Мы исследовали данные за период с 1999 до середины 2019 года и обнаружили 51 120 статей, в которых прямо упоминалось использование одной или нескольких основных библиографических баз данных. Мы предполагаем, что такое признание подразумевает, что для подготовки и публикации научной статьи необходимо проводить гораздо

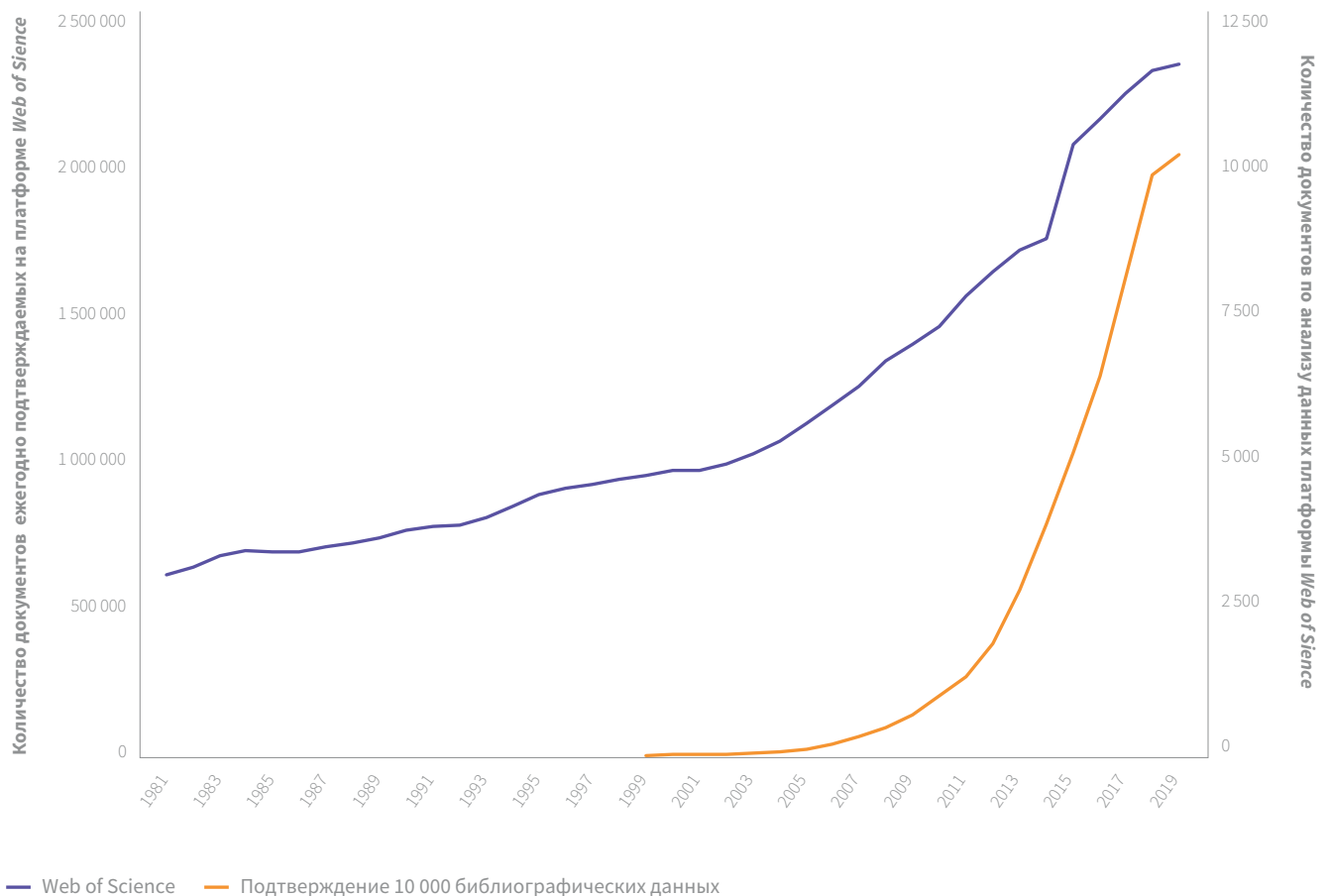
больше поисков в базах научного цитирования, чем принято считать.

Эти данные демонстрируют, насколько быстрым был рост количества таких исследований, по сравнению с общим ростом базы данных, на основе которой они сделаны. И хотя в 1990-е гг в библиографических исследованиях редко прямо ссылались на конкретные базы данных, к 2010 г исследования, в которых такие упоминания есть, увеличивались на 1000 ежегодно, а на данный момент эти цифры возросли многократно. В настоящее время ежегодно публикуется более 10 000 статей, в которых в той или иной форме для обзоров и аналитики используются библиографические данные (Рисунок 1).

Данные демонстрируют, насколько быстрым был рост количества таких исследований, по сравнению с общим ростом базы данных, на основе которой они сделаны.

Рисунок 1

Годовое количество научных документов (статей и обзоров), индексируемых в *Web of Science* с 1984 г, и количество документов, в которых сообщается, что их анализ был сделан на основе источников библиографических баз данных. Данные за 2019 год охватывают только неполный год (до сентября).



Web of Science — не единственный источник данных для такого анализа, ей предшествовали Index Medicus (1879 год) и Biosis (1926 год), но Science Citation Index (непосредственная предшественница *Web of Science*), несомненно, является самой старой базой данных научного цитирования. Существуют и другие важные источники. К ним относится база данных Scopus, созданная в 2004 году издательством Elsevier, а также база данных Google Scholar, которая не занимается тщательным отбором и индексацией публикаций. Для сопоставимости анализа баз данных по дисциплинам мы не рассматриваем в рамках данного отчёта специализированные базы данных PubMed, JSTOR и DBLP, а также базу препринтов arXiv.

Во многих предыдущих исследованиях имеются сведения об использовании более чем одного из этих источников. Пересекающиеся круги диаграммы на рисунке 2 показывают, что из 29 079 публикаций, в которых цитировались данные публикаций *Web of Science*, данные базы Scopus использовались в 7212 (25 %), а данные Google Scholar — в 1488 (5 %). С другой стороны, в 20 511 исследованиях, в которых использовались данные

Scopus, в 35 % случаев также использовались данные *Web of Science* (Рисунок 2).

Это интересный результат. Использование Google Scholar в библиографических целях проблематично, поскольку аналитикам и другим пользователям, проводящим исследования, приходится устранять дублирование и проводить дизамбигуацию информации. Информация о полноте охвата отсутствует. Преимуществом Google Scholar является бесплатный доступ. *Web of Science* и Scopus являются коммерческими базами данных, тщательно отбирающими информацию.

В каких типах публикаций больше всего используются Web of Science?

Большую часть составляют обзоры, но база данных также пользуется популярностью у тех, кто исследует более обычные научные статьи. Использование данных варьируется между этими двумя типами публикаций.

Из числа обзоров около 20 % пользователей выбрали *Web of Science*, а не Scopus, в качестве источника для своего анализа и контента.

Однако, если говорить о статьях, то примерно вдвое больше людей используют данные *Web of Science* по сравнению с данными Scopus. Вероятно, это связано с характером содержания публикации и целями аналитика, которые будут обсуждаться позже в рамках этого отчета (Рисунок 3).

Какова география публикаций?

С давних пор значительная часть авторов приходится на США и Западную Европу, а также англоязычную диаспору Австралии и Канады, но большинство авторов, безусловно, составляют авторы из Китая. В подавляющем большинстве этих случаев основным источником информации для всех этих авторов является *Web of Science*. Однако в Индии и Южной Африке большинство авторов пользуются Google Scholar. А Scopus пользуется большим признанием в Иране и Италии (Рисунок 4).

Рисунок 2.

Количество исследований, опубликованных в научных журналах, проиндексированных в *Web of Science* (1999–2019 гг.), в которых сообщалось об использовании в них одной или нескольких из трех основных баз данных научного цитирования в качестве источника данных для анализа. Данные за 2019 год охватывают только неполный год (до сентября)

- Web of Science 29 079**
- Scopus 20 511**
- Google Scholar 14 079**

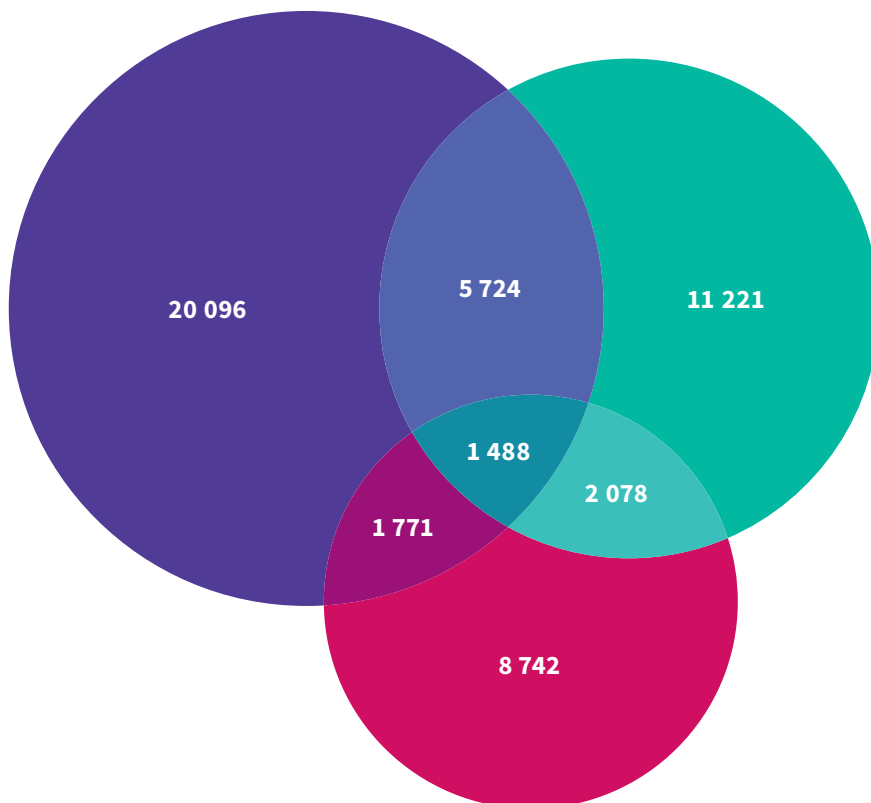


Рисунок 3.

Типы публикаций в научных журналах, проиндексированных в *Web of Science* (1999–2019 гг.), в которых сообщалось об использовании в них одной или нескольких из трех основных баз данных научного цитирования.

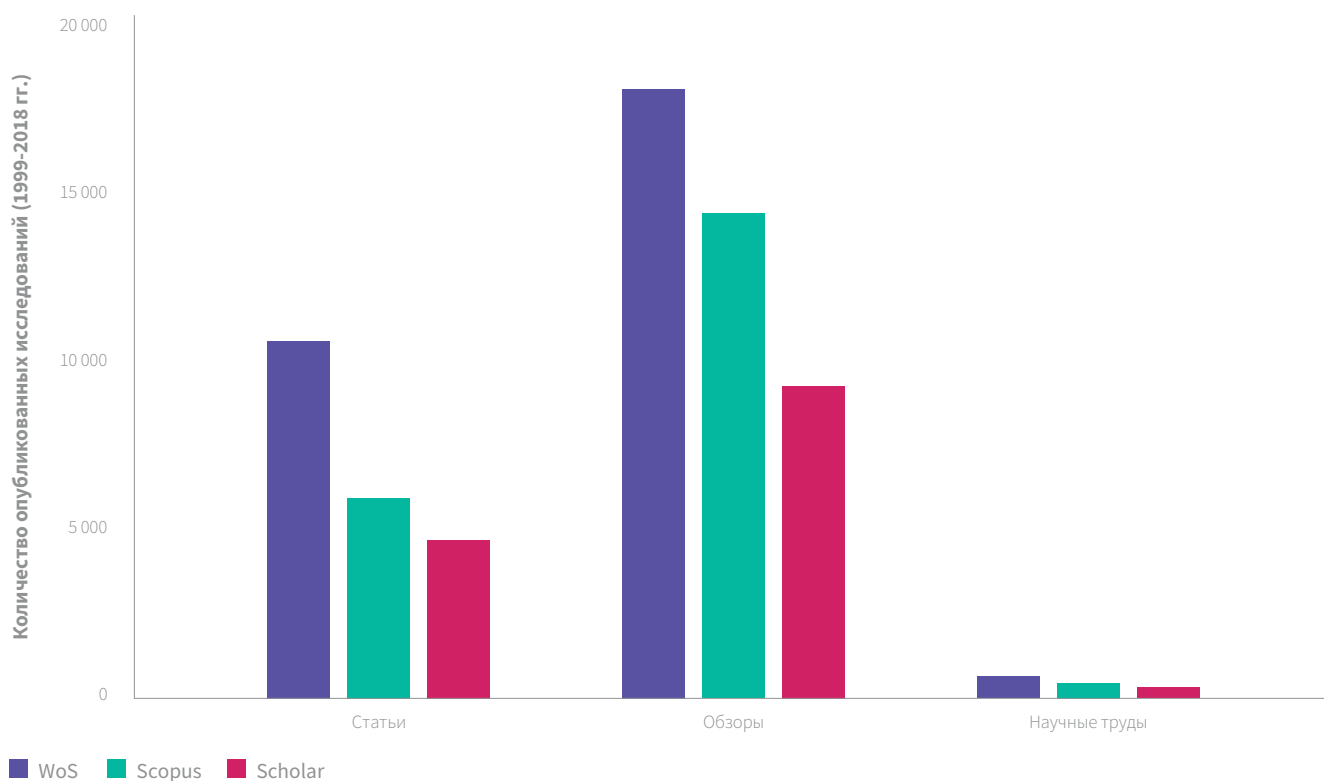
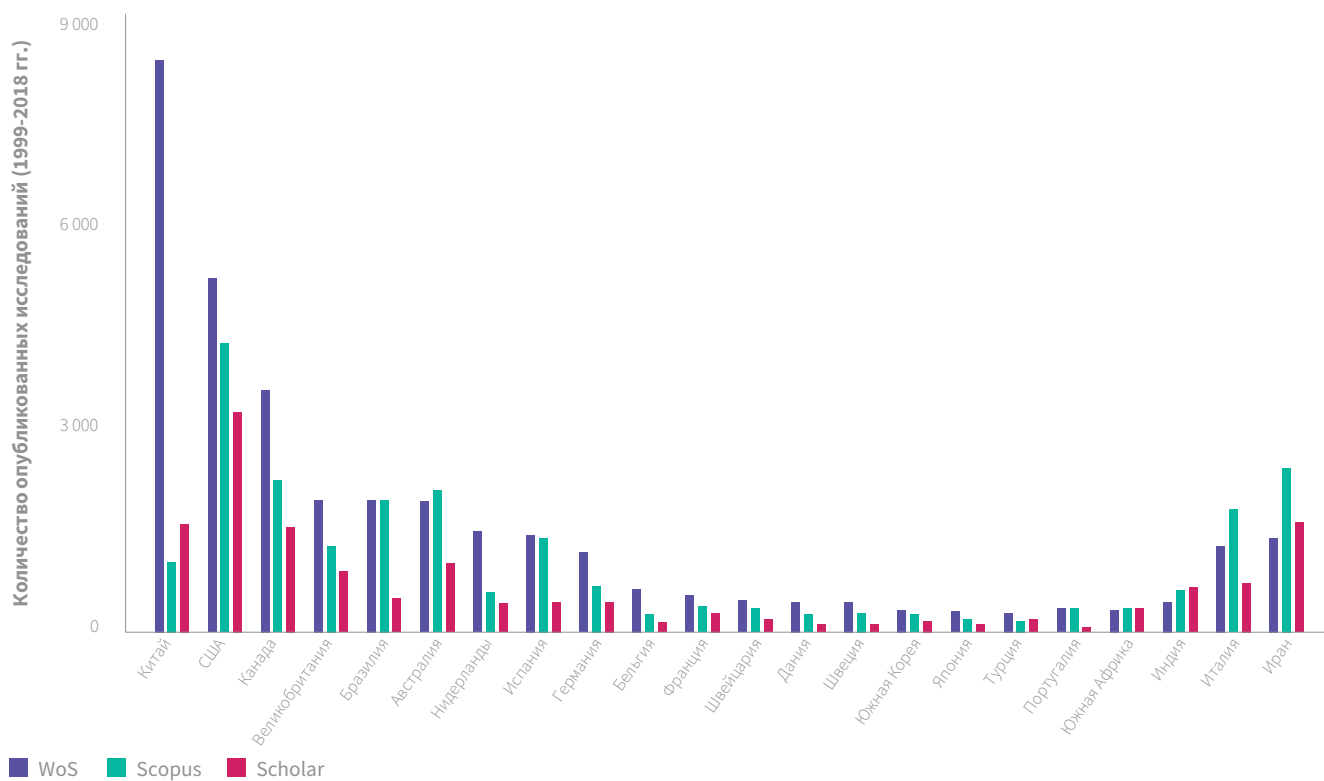


Рисунок 4.

Региональное распределение публикаций в научных журналах, проиндексированных в *Web of Science* (1999–2019 гг.), в которых сообщалось об использовании в них одной или нескольких из трех основных баз данных научного цитирования.



Какие темы включены?

Рост числа публикаций, в которых упоминаются данные *Web of Science* и других библиографических баз данных в качестве основного источника, говорит нам о том, что польза библиографических данных для анализа, а также для поиска и обнаружения, стала очевидной для многих исследователей.

Однако стоит признать, что рост числа таких упоминаний может также отражать изменения в поведении исследователя. На протяжении многих лет учёные использовали данные *Web of Science* в качестве основного и естественного источника для любого библиографического анализа, но с момента появления Google Scholar (выпуска бета-версии в 2004 г), а затем Scopus (2004 г), они признали необходимость включения различных источников, часто для того, чтобы убедительно доказать происхождение и достоверность своей информации.

Каковы характеристики библиографических записей, обеспечивающих эти преимущества? Выделяются два основных варианта использования: систематический обзор и оценка исследований. К обоим вариантам существуют разные подходы.

Как и в любом исследовании в рамках любой области первостепенное значение имеет идентификация и описание используемых источников данных.

Обзоры

Существует множество подходов к обзорам имеющихся библиографических материалов, как правило, для выбранной темы в рамках какой-либо дисциплины. Долгое время обзоры считались одной из наиболее важных частей всего исследовательского корпуса. В связи с этим некоторые обзорные журналы являются одними из наиболее часто цитируемых изданий в своей области. Юджин Гарфилд, основатель ISI, признал важную роль обзоров и их авторов в научной литературе. Он отметил, что утверждения в обзорах были полезны для индексации, и что на каждое утверждение была ссылка, так что цитаты, в более широком смысле, также индексируют «утверждения» (Гарфилд, 1976, 1982 годы).

Некоторые обзоры включают в себя аннотированный список и точные сведения о последних библиографических материалах, в то время как другие — действительно систематичны. В хорошем обзоре суммируются результаты исследований, уже отобранных, принятых и опубликованных в корпусе, а также обеспечивается высокий уровень доказательств результатов предыдущей работы и состояния существующих знаний. Он должен включать в себя суждения о доказательствах и рекомендации для дальнейшей работы. Такой сложный обзор может объединять данные посредством метаанализа с целью обеспечения лучшего качества, поскольку авторы выполняют оценку аргументов за и против.

Для обеспечения такого значительного вклада нет лучшей отправной точки, чем тщательно отобранная, глубоко структурированная библиографическая база данных, содержащая не только большую часть материала, необходимого для обзора, но также предлагающая индексацию и улучшение качества данных, что позволяет автору обзора быстро фильтровать, тщательно исследовать и расставлять приоритеты в рамках рассматриваемого материала. На практике трудно понять, как можно выполнить эффективный

обзор без соответствующей всеобъемлющей базы данных. Автор обзора располагает целым рядом инструментов поиска и должен быть уверен в том, что обнаруженные публикации поступают из авторитетных источников. Эти материалы поступают из журналов, которые должны соответствовать строгим редакционным стандартам. Поэтому неудивительно, что *Web of Science* наиболее часто используется в качестве источника обзорных публикаций.

Оценка исследования

В 1955 году Гарфилд отметил в «Science», что количество цитирований может отражать степень влияния публикации на других исследователей (Гарфилд, 1955 г.). С тех пор область наукометрии значительно расширилась, и многие научные группы в Северной Америке, Европе и, в последнее время, в Австралии и Азии, теперь вносят свой вклад в ее развитие посредством изучения данных и разработки сложных описательных и сравнительных показателей. Эти методы были восприняты руководителями исследований и разработчиками научных стратегий для использования в национальных отчетах и в документах по оценке программ финансирования, где их применимость наиболее эффективна.

Как и в любом исследовании в рамках любой области первостепенное значение имеет идентификация и описание используемых источников данных. Результатом этого станет появление большого количества статей в специализированных по дисциплинам журналах, а также журналах, относящихся к другим дисциплинам, где предметом интереса являются стоимость и преимущества оценки исследований. Из-за увеличения инвестиций в научные исследования во всем мире политическое внимание, уделяемое методам и сравнительным результатам, означает, что данная область получила широкое распространение в региональном анализе.

Методология

Чтобы отобрать темы для анализа, мы изучили текст заголовков, аннотаций и ключевых слов в 51 120 публикаций, извлеченных из *Web of Science* за период с 1999 по середину 2019 года. Использовался стандартный алгоритм моделирования тем Unigram, отфильтровывающий термины, встречающиеся в более, чем 50 % документов, и термины, встречающиеся менее, чем в трех публикациях. Все термины были преобразованы с учетом использования нижнего регистра, в результате получился словарь, содержащий 36 095 слов. Анализ каждого слова (то есть, сколько раз оно встречалось в заголовке, аннотации и ключевых словах) использовался в качестве входных данных для алгоритма неотрицательной матричной факторизации (NMF) с целью создания моделей тем для определенного числа тем. Процесс выполнялся на языке программирования Python с использованием стандартных пакетов библиотеки Scikit learn и бесплатной библиотеки машинного обучения ПО для Python (Pedregosa et al., 2011).

Мы сгенерировали ряд моделей, каждая из которых дала различное количество тем (в диапазоне от 10 до 50), и оценили их качественно, проверив согласованность результатов (то есть, содержала ли модель темы со словами, которые явно связаны между собой?) и гранулярность (то есть, насколько они конкретны в терминологии?). Несмотря на то, что существует возможность создавать тематические модели с очень небольшим или очень большим количеством тем, выбор разумных вариантов для конкретного анализа, как правило, зависит от природы корпуса текстов. Мы обнаружили, что целевой набор тем, которые дали полезный результат, в том смысле, что кластеры имели сравнительно одинаковый размер, содержали четко связанные термины и были достаточно специфичны для разделения набора данных публикации на идентифицируемые кластеры.

На этапе разработки мы не стремились проверить относительное разнообразие тематического контента или степень, в которой они опирались на небольшое количество или множество категорий журналов *Web of Science*. Мы также избегали поиска баланса между годами в рамках каждой

темы. На практике на каждую тему, вероятно, будет оказывать сильное влияние содержание гораздо более многочисленных записей данных за последние несколько лет (см. Рисунок 1). Мы обнаружили, что на основе целевого набора из приблизительно 30 тем удалось создать относительно сбалансированный набор кластеров. Тема — это произвольный раздел в записях данных с использованием порогов сходства. В этом случае мы использовали частоту, с которой термины делятся. Сами темы разделяют эти термины между собой, поэтому мы можем создать семейное древо, которое бы связывало темы друг с другом. Это называется древовидной схемой, которая последовательно объединяет темы в пары, группы и семьи. Темы, объединяемые воедино, более схожи по терминологии, чем темы, которые сильно отличаются друг от друга в древовидной схеме.

Мы обнаружили, что на основе целевого набора из приблизительно 30 тем удалось создать относительно сбалансированный набор кластеров.

На рисунках 5 и 6 показана взаимосвязь между определенными нами темами. У каждой темы есть метка, определяемая на основе анализа наиболее часто используемых терминов в наборе статей по этой теме. Маркировка носит ориентировочный характер, и для правильной интерпретации необходим осознанный анализ фактических работ, которые составляют тему. Маркировка никогда не может быть абсолютной: мнения экспертов расходятся в отношении точного характера материала в любой группе, созданной тем или иным образом; такие взгляды также имеют тенденцию развиваться по мере проведения анализа.

Темы, заслуживающие интереса

Некоторые темы непосредственно связаны с текущими стратегическими приоритетами в области здравоохранения, другие — имеют отношение к методологии, а третьи — указывают на лежащее в основе исследования. Некоторые из них сложно интерпретировать: темы 2 и 5, очевидно, касаются генетики, в частности, полиморфизма и его отношения к лекарственным средствам, специфичным для конкретного пациента, но чтобы понять, почему алгоритм распознает две темы, а не одну, требуется специалист. Тема 26 относится к социальным сетям, но ее следует рассматривать в контексте средств массовой информации как источник информации и дезинформации, особенно в части здоровья.

Доминирующие позиции занимают работы (в основном обзоры), относящиеся к клинической медицине и здоровью, а также к областям исследований в сфере здравоохранения, включая медицинское обслуживание, лечение и инновационные исследования. Это неудивительно. Концепция «систематического обзора» хорошо зарекомендовала себя в биомедицинских исследованиях. Во-первых, поскольку результаты таких исследований являются жизненно важными, передовая практика — ключевое требование для обеспечения быстрого и эффективного прогресса, и этот уровень знаний постоянно пересматривается. Во-вторых, так как эти области являются объектом инвестирования очень большой доли государственных и коммерческих исследовательских фондов, как управление текущими исследованиями, так и будущие инвестиции зависят от регулярного и эффективного мониторинга результатов.

Кокрановская электронная библиотека систематических обзоров (<https://www.cochrane.org/about-us>) особенно важна в этом отношении и оказала глобальное влияние на установление стандартов и демонстрацию ценности и использования. Было изучено использование нескольких баз данных для оказания влияния на такие обзоры. Выводы Bramer et al (2017): «Чтобы гарантировать адекватный и эффективный охват», оптимальные

поиски в систематических обзорах необходимо производить как минимум в Embase, MEDLINE, *Web of Science* и Google Scholar.»

Верхняя часть дерева охватывает международные медицинские приоритеты (диабет, ожирение, дети и беременность), борьбу с болезнями и диагностику (тема 19 указывает на растущую значимость фармакологических стратегий, связанных с растениями) и ряд тем, связанных со все более важными областями психического здоровья и управления. С этим связаны три темы (3, 21 и 22), которые касаются аспектов укрепления здоровья и управления. Исследования в области здравоохранения, а также данные и анализ образуют отдельный кластер (темы 0, 7, 24 и 11). Рядом с основным блоком тем, касающихся медицины и здравоохранения, расположен отдельный блок исследований на ранних этапах в области генетики и молекулярной биологии, связанных с будущими решениями в области здравоохранения.

В нижней части схемы показана довольно четкая группа тем, связанных с экосистемой образования и исследований. Тема 26 относится к социальным сетям и научной коммуникации, а тема 18 — к обучению и навыкам в контексте медицинского обучения. Наконец, публикации, в которых описываются использование данных *Web of Science* для управления исследованиями и инновациями (Тема 1) и оценки исследований (Темы 12 и 27), являются крупнейшими отдельными категориями, в каждой из которых имеется более 2000 связанных статей. Эти три темы также отличаются от остальной части набора данных, потому что в них работы включают в себя в основном статьи, а не обзоры. В целом, они составляют лишь меньшую часть сценариев использования данных. Это может быть сюрпризом для исследователей, считающих, что степень их ответственности и контроль в рамках исследовательской деятельности постоянно увеличиваются. Конечно, эти статьи принадлежат ученым, занимающимся наукометрическими исследованиями. В них описываются тематические исследования, развивающие или иллюстрирующие аналитическую методологию, в отличие от тех, в которых проводятся оценки и которые публикуются реже.

Карта тем и документов

Мы также можем визуализировать темы на карте публикаций, а не в виде семейного дерева на рисунке 5. Карта включает в себя каждую публикацию в нашем наборе данных как отдельную точку, группируя их по текстовому сходству и отображая отношения между темами и блоками, схематически изображенными на древовидной схеме. Это дает нам дальнейшее понимание взаимосвязи между этими группами документов.

После размещения тем среди точек публикации на карте можно добавить другую информацию, чтобы выявить дополнительные шаблоны, которые упростят интерпретацию. В этом случае мы добавили вторую версию карты с разными цветовыми шаблонами публикаций, чтобы идентифицировать те, которые являются информативными (окрашены в красный цвет, в них заголовки или рефераты содержат ключевые слова, такие как «наукометрический», «библиометрический», «информационный») или представляют собой обзоры (выделены зеленым цветом, в их заголовках или аннотациях содержатся такие фразы, как «систематический обзор», «метаанализ» и «поиск литературы»). Этот метод позволяет сразу определить разницу между наукометрическими темами и остальными библиографическими материалами. В результате классификации в общей сложности 4852 научные работы были признаны информативными, а 36 957 получили статус обзоров.

Интересно отметить, что 435 работ были классифицированы и как информметрические, и как обзоры. Это указывает на растущую тенденцию применения наукометрических методов для улучшения методологии систематического обзора, например, для выявления новых областей исследований в рамках дисциплин, для проведения тематического анализа библиографических материалов с использованием графиков цитирования или анализа ключевых слов, а также для изучения тенденций в области финансирования. Они позволяют улучшить понимание принципов действия исследования, предоставляя полезные количественные показатели для оказания влияния на стратегии исследования.

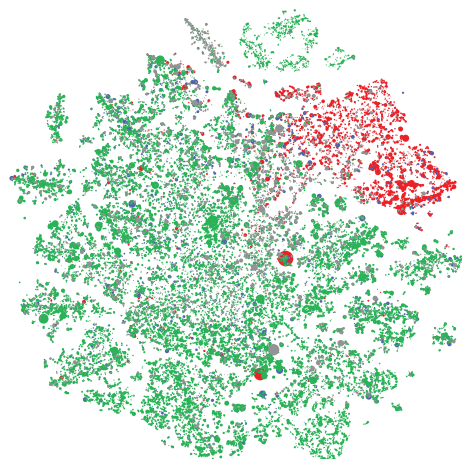
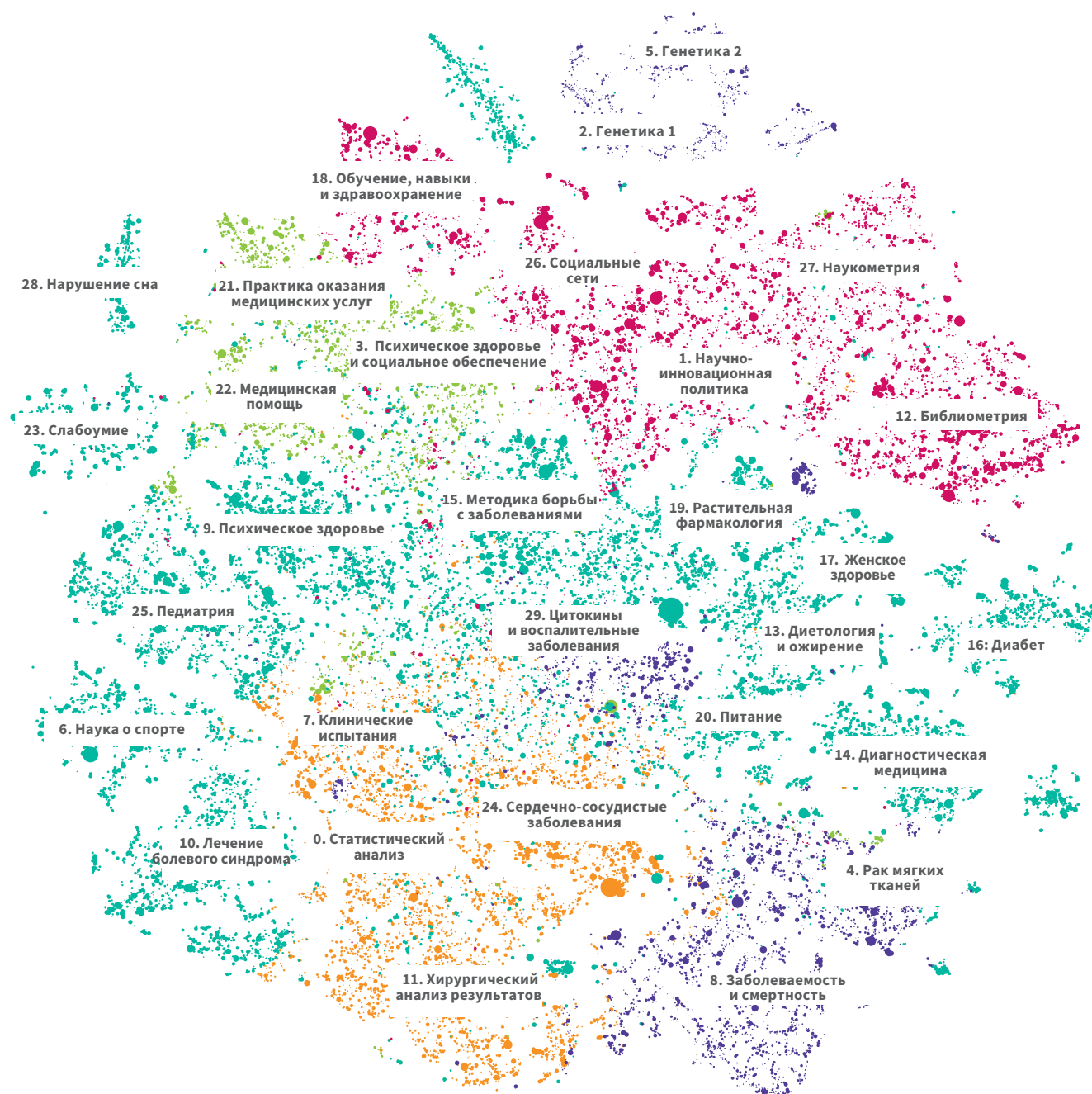
Рисунок 5.

Древовидная схема, показывающая сходство тем, определенных среди 51 120 работ в *Web of Science*, в которых используется библиографическая база данных (публикации за 1999–2019 гг.). Номера статей в каждой теме показаны справа от метки темы. Цветами обозначены темы, которые объединяются вместе с относительно высокой степенью сходства.



Рисунок 6.

Тематическая карта из 51 120 статей (1999–2019 гг.), в которых признается конкретное использование источника библиографических данных.



Комментарий:

На карте показана структура и состав различных тем исследований, в которых используются наукометрические базы данных. Отдельные темы сгруппированы по цветам, соответствующим темам, приведенным на древовидной схеме (см. Рисунок 5). В кластер темно-красного цвета включены темы 18, 26, 1, 27 и 12: темы, в которых использовались методы численного подхода и статистического анализа. Зеленая зона (темы 21, 3 и 22) состоит из тем, которые относятся к социальному обеспечению. Они часто ориентированы на практикующих врачей и близки к темам, затрагивающим конкретные заболевания, которые формируют этот более крупный кластер бирюзового цвета. Разделенный на несколько частей кластер фиолетового цвета (темы 2, 5, 4 и 8) распределен по разным областям, причем в связи с использованием аналитических методологий генетические темы расположены ближе к кластеру темно-красного цвета.

Рост числа тем и их распределение по категориям

Рост числа определенных тем с течением времени в целом отражает общую картину, показанную на рисунке 1. Количество тем заметно возрастает в период после 2008 года, в последние же несколько лет рост усилился в значительной степени.

Единственными отличающимися темами (только по степени, а не по основному профилю) являются наукометрические темы (12 и 27), исследователи которых стали чаще упоминать использование конкретных баз данных раньше, чем в других темах. После чего последующая траектория роста была линейной, а не следовала за движением экспоненциального профиля, наблюдаемого в биомедицинских темах.

Темы, созданные из текста в наборе данных из 51 120 названий статей, рефератов и ключевых слов, представляют собой общее представление об этой деятельности. Эти документы были опубликованы в самых разных журналах, и любая отдельная тема может состоять из компонентов не только из областей дисциплины, что ожидаемо для специалиста

в данной области, но также из других областей, связанных с поддержкой исследований и технологий.

Существует 254 категории журналов *Web of Science*, но более половины из 51 120 статей, рассмотренных в этом отчете, были опубликованы только в 27 из них (то есть в 11 % категорий журналов).

На долю учетных записей Information Science приходится 1378 (2,7 %) от общего количества работ, подтверждающих использование в них библиографической информации. Информатика (889) и исследования в области образования (476) также относятся к числу наиболее часто запрашиваемых категорий. Остальные категории имеют биомедицинскую и ориентированную на здоровье тематику, в верхней части таблицы расположены категории Онкология (2104) и Общественное здравоохранение (2020) (таблица 1).

Существует ожидаемая схема включения между конкретными категориями *Web of Science* и индикативной маркировкой тем. Например, тема 24 была определена как связанная главным образом

с управлением рисками при сердечно-сосудистых заболеваниях, и неудивительно, что 572 (28 %) из 2008 статей, связанных с этой темой, можно найти в журналах в категории *Web of Science* «Сердечная и сердечно-сосудистая системы», что составляет почти половину всех кардиологических журнальных работ (см. таблицу 1), поэтому остальные статьи относятся к другим темам и распределены среди них.

Другие три четверти статей, связанных с темой 24, были взяты из огромного разнообразия журналов. Хотя большинство из них были сгруппированы в ожидаемых областях медицинских и смежных исследований, распространение других категорий журналов, включая химию и инженерию, является показателем все более междисциплинарного характера передовых соответствующих исследований, отраженных в эффективном обзоре. Полнота и точность сети ссылок на цитаты в должным образом отобранной библиографической базе данных становятся неотъемлемой частью искусства для автора тематического обзора. Неявных экспертных знаний уже недостаточно.

Таблица 1.

Категория журнала Web of Science	До 2009 г.	2013–2019	2014	2015	2016	2017	2018	2019 г. (часть)	Общее количество
Общий итог	1 166	7552	3 860	5 047	6 302	7 943	9 654	9 596	51 120
Онкология	28	330	283	205	263	266	370	359	2 104
Общественное здоровье, экологическая и профессиональная медицина	40	286	160	193	237	331	390	383	2 020
Информатика и библиотечное дело	93	323	111	139	133	197	240	142	1 378
Стоматология, хирургическая стоматология и медицина	41	167	88	116	168	203	285	297	1 365
Фармакология и фармацевтическое дело	45	256	92	110	171	188	246	243	1 351
Хирургия	19	162	74	135	186	204	277	281	1 338
Уход за больными	24	162	80	110	159	197	226	252	1 210
Гастроэнтерология и гепатология	30	176	115	138	154	167	181	185	1 146
Эндокринология и обмен веществ	26	162	89	93	118	169	219	254	1 130
Сердечная и сердечно-сосудистая системы	27	164	91	112	132	168	209	161	1,064

Применение в будущем

По мере роста количества фактов исследователи могут повторно обращаться к базе данных научного цитирования *Web of Science* для сравнительного анализа наиболее полезного материала.

Как уже отмечалось ранее, в кластере тем на рисунке 5 преобладают темы медицины/здоровья (13 тем связаны с болезнями, три — с лечением болезней, четыре — с медицинскими данными и пять — с базовыми исследованиями в области биологии). Остальные пять тем охватывают области образования и управления исследованиями.

Большую часть кластерных документов по всем темам, кроме двух в области управления исследованиями, можно классифицировать как обзоры, тогда как большинство документов, в которых признается использование *Web of Science* и других библиографических баз данных, и которые можно классифицировать как статьи, относятся к области управления исследованиями.

Оценка и политика исследований, а также управление, являются ключевым элементом поддержки наукометрических исследований записями публикаций *Web of Science* и соответствующими данными. В этой области данные *Web of Science* являются основным источником информации и используются гораздо чаще, чем любые другие. Юджин Гарфилд был ключевой фигурой в процессах происхождения, ранней эволюции и формирования основополагающих идей наукометрии, а его база данных SCI, предшественница *Web of Science*, является прародительницей этой области. Его наследие продолжает оказывать глобальное влияние на область и в настоящее время.

Обзоры в биомедицинских темах основаны на разнообразии категорий *Web of Science* (Таблица 1). Это напоминание о вкладе, внесенном множеством дисциплин, но чего не хватает в тематической карте, так это какого-либо существенного компонента, связанного с естественными науками, технологиями и гуманитарными науками (Рисунок 6).

Этот результат обусловлен двумя аспектами: во-первых, относительной зрелостью обзора литературы как инструмента, лежащего в основе исследований в области биомедицины; во-вторых, относительной частотой и распространенностью обзоров как части управления биомедицинскими исследованиями. Это обусловлено как масштабом инвестиций, так и политическим приоритетом, лежащим в основе этих инвестиций.

В рамках обзоров в сфере биомедицины не только рассматриваются библиографические материалы, но и производится детализация результатов исследований. Источник данных становится важной частью контрольного журнала, где спрос на лучшие практики также является попыткой оптимизировать результаты в отношении потраченных средств.

Хотя обзоры также прочно утвердились и имеют признанное значение и ценность в других дисциплинах (например, Annual Review of Condensed Matter Physics, Annual Review of Materials Research), они являются скорее вежами, а не частью управления исследованиями. Данные исходных исследований, содержащиеся в *Web of Science*, также усваиваются и передаются по другим маршрутам, часто в мире «серой литературы», которая включает в себя отчеты правительств и агентств, а также профессиональные публикации. Инновационной моделью является периодический обзор на основе Интернета, который используется учеными, изучающими изменение климата, на веб-сайте ScienceBrief (sciencebrief.org/about), где представлен анализ отдельных научных работ в виде «прозрачной, непрерывной и быстрой системы анализа текущих знаний». По мере роста количества фактов исследователи могут повторно обращаться к сети цитирования *Web of Science* для сбора наиболее полезного материала.

Ссылки и информационно-справочные материалы

Bramer, W. M., Rethlefsen, M. L., Kleijnen, J. и Franco, O.H. (2017). Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Systematic Reviews*, 6, статья № 245.

Garfield, E. (1955). Citation indexes for science: a new dimension in documentation through association of ideas. *Science*, 122, стр. 108–111.

Garfield, E. (1976). Significant journals of science. *Nature*, 264, стр. 609–615.

Garfield, E. (1982). ISI's "new" Index to *Scientific Reviews* (ISR): applying research front specialty searching to the retrieval of the review literature. *Current Contents*, 39, стр. 5–12.

Garfield, E. (1987a). Reviewing review literature: Part 1 — definitions and uses of review. *Current Contents*, 18, стр. 5–8.

Garfield, E. (1987b). Reviewing review literature: Part 2 — the place of reviews in the scientific literature. *Current Contents*, 19, стр. 5–10

Gurney, K. A. and Pendlebury, D. (2012). *Funding acknowledgements in the journal literature: uses and benefits for funders, recipients, and analysts*. Philadelphia, PA: Thomson Reuters.

Li, K., Rollins, J. и Yan, E. (2018). *Web of Science* use in published research and review papers 1997–2017: a selective, dynamic, cross-domain, content-based analysis. *Scientometrics*, 115, стр. 1–20.

Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss и V. Dubourg. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, стр. 2825–2830.

Pringle, J. (2008). Trends in the use of ISI citation databases for evaluation. *Learned Publishing*, 21, стр. 85–9.

Schnell, J. D. (2018). *Web of Science: the first citation index for data analytics and scientometrics*. In *Research Analytics: Boosting University Productivity and Competitiveness through Scientometrics*. Под редакцией Francisco J. Cantu-Ortiz. Boca Raton, FL: CRC Press, стр. 15–29. ISBN-13: 978-1498785426

Сведения о серии международных научно-исследовательских отчетов Института научной информации (ISI)

В основе наших международных научно-исследовательских отчетов лежат проведенные нами уникальные отраслевые исследования. Их целью являются выводы, анализ, идеи и комментарии для просвещения и стимулирования дискуссий. В каждой из этих серий отчетов демонстрируется огромный потенциал данных исследований для освещения проблем управления в области оценки исследований и политики исследований и для ускорения развития международной исследовательской базы.

Ссылка для загрузки:

www.webofsciencegroup.com/isi

Сведения о Web of Science Group

Web of Science Group в лице компании Clarivate Analytics занимается сбором научной информации по всему миру для того, чтобы предоставить ученым, компаниям, издателям и правительствам возможность ускорить темпы развития исследований. Ключевым элементом является база данных *Web of Science* — крупнейшая в мире издательско-нейтральная база научного цитирования и исследовательская аналитическая платформа. В число многочисленных известных брендов компании входят *Converis*, *EndNote*, *Kopernio*, *Publons*, *ScholarOne* и Институт научной информации (ISI).

«Университет» Web of Science Group, ISI, поддерживает корпус знаний, на которых строится эта база данных и связанный с ней информационный и аналитический

контент и услуги, и распространяет эти знания посредством мероприятий, конференций и публикаций, а также проводит исследования для поддержания, расширения и совершенствования базы знаний. В ISI постоянно ведутся работы по изучению существующих стандартов. Институт сотрудничает с партнерами из государственного и частного секторов с целью изучения данных публикаций и цитирования, а также с целью создания инновационных индикаторов ценности для исследователей и руководителей исследований.

Для получения дополнительных сведений посетите сайт webofsciencegroup.com

эл. почта: ISI@clarivate.com



[@webofscience](https://twitter.com/webofscience)

webofsciencegroup.com/isi

© Clarivate Analytics 2019. ISI, Web of Science Group и ее логотип, а также все другие товарные знаки, используемые в данном документе, являются товарными знаками соответствующих владельцев и используются по лицензии.

WS445086582/01