И.Н. ЧЕРНОВА,
ФГБНУ «Дирекция НТП» (Москва, Российская Федерация; e-mail: chernova@fcntp.ru)
О.В. ЧЕРЧЕНКО,
ФГБНУ «Дирекция НТП» (Москва, Российская Федерация; e-mail: olya.cherchenko@mail.ru)

АЛГОРИТМ УТОЧНЕНИЯ РАМОК НАУЧНО-ТЕХНОЛОГИЧЕСКОЙ ОБЛАСТИ В БИБЛИОМЕТРИЧЕСКИХ БАЗАХ ДАННЫХ НА ПРИМЕРЕ СИНХРОТРОННЫХ, НЕЙТРОННЫХ ИССЛЕДОВАНИЙ И РАЗРАБОТОК

УДК: 001
10.22394/2410-132X-2022-8-2-98-117

Аннотация: Для оценки публикационной результативности Федеральной научно-технической программы развития синхротронных и нейтронных исследований и исследовательской инфраструктуры на 2019–2027 гг. (Программа) ее разработчиками предложен специальный поисковый запрос в Web of Science Core Collection (WoS). Цель исследования – определение эффективных подходов к доработке поискового запроса, обеспечивающих наиболее полный охват публикаций по темам Программы и применимых для международных и отечественных баз данных. В публикациях из заявок на участие в Программе выявлены ключевые слова для валидации поискового запроса. Использован поиск по ключевым словам в WoS, Scopus, ядре Российского индекса научного цитирования, идентификация публикаций с помощью API WoS и Scopus, построение графов в VOSviewer, метод текст-майнинга Yet Another Keyword Extractor. На основании эмпирических данных предложен многоступенчатый алгоритм формирования коллекции публикаций конкретной научно-технологической области в библиометрических базах данных.

Ключевые слова: *индикаторы, публикационная активность, поисковый запрос, библиометрический анализ, текст-майнинг, синхротронные исследования, нейтронные исследования*

Благодарность: Работа выполнена при финансовой поддержке Минобрнауки России в рамках соглашения о предоставлении субсидии № 075-02-2022-979 от 22.02.2022 г.

Для цитирования: Чернова И.Н., Черченко О.В. Алгоритм уточнения рамок научно-технологической области в библиометрических базах данных на примере синхротронных, нейтронных исследований и разработок. *Экономика науки*. 2022; 8(2):98–117. https://doi.org/10.22394/2410-132X-2022-8-2-98-117

ВВЕДЕНИЕ

В условиях конкурентной исследовательской среды публикация результатов научного труда служит важным свидетельством достижений исследований и разработок и играет значительную роль в накоплении сравнительных преимуществ ученых, организаций, а также в усилении позиций научного сообщества в конкретных областях науки. Этим объясняется применение библиометрического показателя «число/количество публикаций в высокорейтинговых научных журналах» (и/или «число/количество публикаций, индексируемых в российских и международных информационно-аналитических системах научного цитирования») для решения широкого круга задач современной научной политики России: научного и информационного сопровождения принятия решений о поддержке исследований; систематического анализа развития различных областей науки; мониторинга и изучения результативности научной

деятельности организаций, научных коллективов, авторов и т.д.

Оценка эффективности хода реализации федеральных программ и проектов в России, в том числе на основании расчета фактически достигнутого значения целевого индикатора (показателя) количество публикаций, также входит в число задач, на которые направлено внимание заказчиков-координаторов, ответственных исполнителей программ в научно-технической сфере, например, при принятии решений о результативности распределения ресурсов, корректировке программ поддержки исследований и разработок, определении перспективных направлений. Вместе с тем, как показывают исследования, количественные показатели публикационной активности являются недостаточными для мониторинга научной деятельности и формирования корректных выводов и оценок, вызывая различные искажения результатов, формализацию оценок и критику со стороны научного сообщества, а значит требуют компетентного и аккуратного использования [1–4].

Сфера применения библиометрического показателя «количество публикаций» охватывает оценку результативности научных исследований и разработок на разных уровнях агрегирования информации о публикациях – авторов и их институциональной принадлежности, регионов и стран местонахождения, областей науки и техники, источников финансирования и др. [5]. Этот количественный показатель обычно применяется на обширной совокупности данных о публикациях путем поиска и ранжирования необходимых объектов оценки. При этом одной из важных проблем является идентификация публикаций в определенной научно-технологической области, выделение специализированных научных результатов в междисциплинарной научно-исследовательской и информационной среде.

В области наукометрии методы классификации научных публикаций занимают особое место, поскольку это влияет на выявление традиционных и новаторских тематик исследований, на методы анализа научной деятельности и расчет показателей ее результативности и даже на то, как создаются новые научные организации [6]. В настоящей статье рассматривается способ выделения и уточнения рамок научно-технологической области на примере синхротронных и нейтронных исследований и разработок, поддержке которых посвящена Федеральная научно-техническая программа развития синхротронных и нейтронных исследований и исследовательской инфраструктуры на 2019–2027 гг. (далее – Программа). На основании эмпирических данных в статье предложен многоступенчатый алгоритм формирования коллекции публикаций конкретной научно-технологической области в библиометрических базах данных. Особое внимание уделено анализу терминологических и синтаксических рамок поискового запроса для получения объективных сведений о количестве публикаций российских авторов по научно-технологическим направлениям реализации Программы.

ОБЗОР ЛИТЕРАТУРЫ

Выделяют три основных способа распределения публикаций по областям исследований, которые подробно описаны в [6]: соотнесение публикаций с категориями научных журналов; алгоритмическая классификация на уровне публикации; контролируемый поиск информации – объединение статей в одну группу по ключевым словам, названиям журналов и другим параметрам.

Первый способ, опирающийся на классификаторы журналов, наиболее прост и доступен для автоматического поиска публикаций по заданным областям. Такой способ позволяет извлечь из реферативных, библиографических и библиометрических баз данных коллекции публикаций в соответствии с классификационными системами и преимущественно применяется в наукометрических исследованиях [6–8]. Одной из основных проблем в работе с классификаторами журналов является неточность соотнесения по журнальным группам статей, особенно если они опубликованы в междисциплинарных журналах [9], журналах из других научных областей [10] или созданы в узких научных областях, категории по которым отсутствуют в используемых классификаторах [9].

Второй способ – алгоритмическая классификация – использует взаимосвязи между публикациями, а не журнальные категории. Распределение на группы классификаторов здесь часто основано на анализе пересечений пристатейной библиографии, со-цитирований [6], методах искусственного интеллекта или других подходах [11]. Способ применяется для создания карт научных областей и в таких решениях, как Топики цитирования (Citation Topics, представлены в InCites), классификационная система The Fields of Research (FOR на платформе Dimensions), охватывающая не только публикации в научных журналах, но и патентные документы, отчеты о клинических испытаниях и другие документы. Недостатками классификации на уровне публикаций является частая непрозрачность используемых алгоритмов, неточность подбора групп классификаторов [12], которая может быть решена с помощью привлечения экспертов [13] или методами машинного обучения [9].

Третий способ – контролируемый поиск информации, по меткому свидетельству специалистов Института научной информации, повышает вероятность подбора гомогенных публикаций, однако имеет существенный недостаток: не многие обладают необходимым опытом в сфере поиска информации в анализируемой области, чтобы создавать специализированные наборы статей, кроме того, эта работа отличается низкой воспроизводимостью [6].

ИНДИКАТОР ПУБЛИКАЦИОННОЙ РЕЗУЛЬТАТИВНОСТИ ПРОГРАММЫ

Способ контролируемого поиска публикаций был взят на вооружение разработчиками Программы. Методика расчета значений целевого индикатора «Количество публикаций в области синхротронных и нейтронных исследований (разработок) в журналах, индексированных в международных базах данных» (далее – Методика, индикатор И1) регламентирует в пункте 7 условия поиска научных статей российских авторов в базе данных «Сеть науки» (Web of Science Core Collection) в виде поискового запроса по темам «импульсный источник нейтронов на основе реакции испарительно-скалывающего типа», «рассеяние нейтронов», «дифракция нейтронов», «синхротрон» [14].

Количество публикаций занимает одну из важных позиций среди целевых индикаторов и показателей Программы, но не имеет первостепенного значения. В числе основных мероприятий Программы – создание и развитие в России «сетевой синхротронной и нейтронной инфраструктуры», в которой исследовательские объекты класса «мегасайенс» рассматриваются «как единая сеть» [15]. В рамках Программы запланированы модернизация Курчатовского специализированного источника синхротронного излучения «КИСИ-Курчатов» (г. Москва), строительство синхротрона поколения 4+ центра коллективного пользования «Сибирский кольцевой источник фотонов» (ЦКП «СКИФ», Новосибирская область) и уникальной научной установки класса «мегасайенс» на о. Русский (г. Владивосток), а также создание прототипа импульсного источника нейтронов на основе реакции испарительно-скалывающего типа (г. Протвино Московской области) [14–17]. Кроме того, Программой предусмотрено создание не менее 25 исследовательских станций Международного центра нейтронных исследований на базе высокопоточного реактора «ПИК» (г. Гатчина Ленинградской области), самого мощного в мире исследовательского реактора [14, 18]. Сетевая инфраструктура, охватывающая значительную территорию страны, представляется перспективной для интенсивного пространственного развития. Подобные объекты как центры концентрации фундаментальных знаний, подготовки высококвалифицированных научных кадров мирового уровня дадут стимул развитию новых продуктов и технологий, созданию высокотехнологичных производств, рабочих мест и привлечению инвестиций [18].

В целом наибольшее внимание в Программе уделено подготовке условий для достижения конкурентоспособных научных и научно-технических результатов, а также высококвалифицированных специалистов для создания синхротронных и нейтронных источников для проведения синхротронных и нейтронных

исследований (разработок), а в качестве приоритетных выступают, например, индикаторы «Количество введенных в эксплуатацию в рамках реализации Программы экспериментальных станций на отечественных синхротронных и нейтронных установках», «Количество адаптированных и разработанных в рамках реализации Программы ускорительных и реакторных технологий, технических решений» и др. [14]. Очевидно, с учетом приоритетности подобных индикаторов для оценки публикационной результативности разработчики Программы не применили в полной мере точного формата поискового запроса, включающего необходимые ограничительные и связующие элементы поиска или иные требования.

В ходе работ по соглашению ФГБНУ «Дирекция НТП» с Минобрнауки России № 075-02-2021-1538 от 22.06.2021 г. авторами статьи была обоснована необходимость коррекции поискового запроса, предложенного в рамках Программы, для проведения объективного поиска публикаций и формирования наиболее полной коллекции научных статей ***в области исследований и разработок, выполненных с использованием синхротронного и нейтронного излучения, а также в области развития ускорительных и реакторных технологий, основанных на использовании синхротронного и нейтронного излучения, опубликованных российскими исследователями*** [14]. Кроме того, подготовлены данные, которые могут быть использованы при усовершенствовании поискового запроса с применением экспертных знаний. Однако проведение дальнейших работ было осложнено не только приостановкой доступа к Web of Science, но также изменениями административного и организационного характера.

В соответствии с постановлением Правительства Российской Федерации от 19.03.2022 г. № 414 «О некоторых вопросах применения правовых актов Правительства Российской Федерации, устанавливающих требования, целевые значения показателей по публикационной активности» до 31.12.2022 г. [19], не применяются требования о наличии публикаций (публикационной активности) в изданиях (научных изданиях), журналах, индексируемых в международных базах данных (информационно-аналитических системах научного цитирования) (Web of Science, Scopus) при осуществлении мер государственной поддержки (предоставлении грантов в форме субсидий) научных, научно-технических проектов, а также при оценке результативности таких проектов. Помимо этого, в Правительство РФ 01.06.2022 г. поступили предложения от рабочих групп по формированию Национальной системы оценки результативности научных исследований и разработок. В этих предложениях предполагается создание «белого списка журналов», который, вероятно, будет состоять из модифицированного ядра Российского индекса научного цитирования (ядро РИНЦ), размещенного на платформе eLIBRARY.RU (https://www.elibrary.ru/).

Ядро РИНЦ представляет собой подмножество статей, опубликованных в журналах, включенных хотя бы в одну из трех баз данных:

- Russian Science Citation Index (RSCI) – отобранные с помощью совокупности библиометрических показателей, формальных критериев, экспертной оценки и общественного обсуждения «лучшие российские журналы» (https://www.elibrary.ru/rsci_about.asp);
- Web of Science Core Collection (WoS);
- Scopus.

В целом даже при отсутствии доступа к таким международным библиометрическим базам данных, как WoS и Scopus, оценка может производиться во многом по тем же журналам, что и раньше, а значит, и тем же данным, отражающим вовлеченность российских исследований в мировую науку. При этом ядро РИНЦ, благодаря данным из RSCI, позволяет учесть публикации в журналах, которые не оказывали сильное влияние на международном уровне, но оценены как сильные в России. В 2022 г. возросла роль таких журналов как площадок для обсуждения научно-технических идей, позволяющих решить локальные задачи, возникшие в России из-за целого ряда экономических, политических и торговых ограничений.

Учитывая происходящие изменения, целесообразно определить наиболее эффективные подходы к доработке поискового запроса

для оценки публикационной результативности в области синхротронных и нейтронных исследований и разработок, обеспечивающие полноту информации и применимые для различных информационных ресурсов: не только международных, но и отечественных баз данных. Предположение о том, что ключевые слова из публикаций участников и победителей конкурса, проводимого Минобрнауки России в рамках Программы, могут быть основой для валидации представленного в Программе поискового запроса, является гипотезой настоящей статьи.

Для проверки гипотезы проанализирован массив публикаций, полученный по результатам конкурса Программы, и сформирован список содержащихся в них ключевых слов; проведено сравнение нескольких вариантов поисковых запросов на основе данного списка с использованием как предусмотренных в рамках Программы терминов, так и новых, дополнительно отобранных ключевых слов в исследуемой научно-технологической области.

МАТЕРИАЛЫ И МЕТОДЫ

Для выполнения исследования использован актуальный на момент его проведения массив публикаций участников и победителей конкурса, реализуемого в рамках Программы. Сведения о публикациях, представленных участниками и победителями конкурса согласно требованиям пп. 1.3–1.4 приложения 2 к форме 4 «Сведения о квалификации» конкурсной документации Программы, получены 12.12.2021–13.12.2021 г. с Портала регистрации заявок на участие в конкурсе (http://prz.sstp.ru) [20]. Всего проанализировано 53 заявки, в том числе 21 заявка победителей, включая данные о наиболее значимых научных публикациях в журналах, индексируемых в WoS и/или Scopus, представляющих результаты исследований с использованием синхротронного и нейтронного излучения за период с 01.01.2016 г.

Полученная выгрузка публикаций прошла предварительную обработку для однозначной идентификации и верификации данных в WoS по полям «Accession Number Web of Science», «Названия основных научных публикаций, подтверждающих квалификацию с 01.01.2016» и «DOI публикации», а в Scopus – по полям «EID (Electronic Identifier) Scopus», «Названия основных научных публикаций, подтверждающих квалификацию с 01.01.2016» и «DOI публикации». По заданным параметрам произведена очистка от лишних символов, исправлены ошибки, препятствующие автоматическому поиску, и выполнена серия выгрузок с использованием API Web of Science – API Expanded уровень Basic и API Lite, API Scopus. В результате идентифицированы 4156 публикаций из изданий, индексируемых в WoS, и 4307 публикаций из изданий, индексируемых в Scopus, авторами которых являются участники конкурса. Победители конкурса являются авторами 2009 (WoS) и 2075 (Scopus) публикаций.

Контролируемый поиск и анализ количества публикаций по ключевым словам в области синхротронных и нейтронных исследований и разработок проведен за 2016–2022 гг. с использованием баз данных WoS, Scopus и ядра РИНЦ на платформе eLIBRARY.RU.

В целях обработки контекстной информации публикаций в работе использовано сочетание методов и техник анализа сетей (network science) и текстов (text mining):

- построение графов и визуализации библиометрических сетей с использованием программного инструмента VOSviewer, позволяющее оценить частоту встречаемости ключевых слов и их взаимодействия;
- извлечение из названий и аннотаций публикаций ключевых слов с помощью метода Yet Another Keyword Extractor (YAKE) [21].

В настоящее время все большее применение в области изучения динамики науки находит анализ текстов и ключевых слов документов, а также взаимосвязей между ними и представление полученных данных на картах науки с использованием вычислительных методов [22]. Текст-майнинг (text mining) или интеллектуальный анализ текстов используют, как правило, в качестве технологии извлечения новой и ценной информации, обнаружения закономерностей в массивах текстовых данных [23]. Его суть заключается в идентификации наиболее часто встречающихся

ключевых слов, которые становятся основой для исследовательских выводов [24]. Несмотря на неспособность алгоритмов самостоятельно определять суть лингвистических понятий, текст-майнинг успешно применяется как вспомогательный метод для решения различных задач [25].

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Как было указано выше, в соответствии с Методикой расчета значений целевых индикаторов и показателей Программы [14], значение индикатора И1 определяется на основе поискового запроса в WoS по темам «импульсный источник нейтронов на основе реакции испарительно-скалывающего типа», «рассеяние нейтронов», «дифракция нейтронов», «синхротрон».

Экспертным сообществом Программы был предложен запрос в WoS, который выглядит следующим образом:

TS = ((spallat* neutron source) OR (neutron scattering) OR (neutron diffraction) OR (synchrotron)) Refined by: DOCUMENT TYPES: (ARTICLE) AND COUNTRIES/REGIONS: (RUSSIA)

Результаты поиска по данному поисковому запросу за 2016–2022 гг. на 29.04.2022 г. представлены в столбце 3 *таблицы 1*.

Если воспроизвести этот же поисковый запрос в доступной на территории Российской Федерации базе данных Scopus, то он будет представлен следующим образом:

TITLE-ABS-KEY ((spallat* AND neutron AND source) OR (neutron AND scattering) OR (neutron AND diffraction) OR (synchrotron)) AND (LIMIT-TO (AFFILCOUNTRY, "Russian Federation")) AND (LIMIT-TO (DOCTYPE, "ar"))

Результаты поиска по данному поисковому запросу за 2016–2022 гг. на 10.06.2022 г. представлены в столбце 4 *таблицы 1*.

Аналогичный поиск выполнен в ядре РИНЦ. На платформе eLIBRARY.RU ядро РИНЦ представлено единой коллекцией из трех информационных ресурсов, исключающей дублирование информации, по сравнению со сбором и обработкой данных по одним и тем же публикациям одновременно в WoS и Scopus, списки индексируемых журналов которых частично перекрываются. Кроме того, доступен поиск не только на английском, но и на русском языке. Однако тематический поиск в eLIBRARY.RU и анализ найденных результатов без применения API имеет ряд сложностей и ограничений по сравнению с международными аналогами. Так, при получении данных по отдельным годам из ядра РИНЦ в стандартном интерфейсе eLIBRARY.RU нет возможности использовать специальные фильтры, построить статистические отчеты или сделать выгрузки. Для каждого года необходимо составлять отдельный поисковый запрос и сохранять данные по каждому из них в отдельную подборку. Использование логических операторов, усечений и операторов близости также имеет целый ряд нигде не прописанных ограничений. Например, при попытке воспроизвести аналогичный представленному выше поисковый запрос с использованием усечения терминов в виде знака астериска (*) и оператора OR (ИЛИ) без ограничения близости расположения терминов поиск приводит к нулевой выдаче. То есть не накладывающий ограничения на расстояние между ключевыми словами поисковый запрос (spallat* neutron source) OR (neutron scattering) выдает ошибку, в то время как поисковый запрос с использованием точных фраз «spallat* neutron source» OR «neutron scattering» выявил более 2000 публикаций.

Для более полного охвата релевантных предложенному поисковому запросу и анализируемой теме публикаций, существующих в ядре РИНЦ, было применено усечение с помощью знака астериска для всех ключевых слов, а также их перевод на русский язык. В итоге был составлен следующий поисковый запрос:

«spallat* neutron* source*» OR «источник* нейтрон* делен*» OR «neutron* scatter*» OR «рассеян* нейтрон*» OR «neutron* diffract*» OR «дифракц* нейтрон*» OR «synchrotron*» OR «синхротрон*» (поиск в названии публикации, в аннотации, в ключевых словах, тип публикации – статьи в журналах, без учета морфологии)

Результаты поиска по этому запросу за 2016–2022 гг. на 14.06.2022 г. представлены в столбце 5 *таблицы 1*. При этом в eLIBRARY.RU осложнен и дальнейший анализ полученных

данных, так как статистические отчеты в подборке формируются для всей коллекции РИНЦ, а не его ядра.

Как предложено экспертным сообществом Программы, поиск производится по названию публикации, аннотации и ключевым словам (в WoS набор данных полей обозначается как TS). Для выявления ключевых слов, которые вероятнее всего могут стать репрезентативной основой поискового запроса в исследуемой научно-технологической области, проанализированы публикации участников и победителей конкурса, основанные на исследованиях, проведенных с использованием синхротронного и нейтронного излучения за период с 01.01.2016 г.

По идентификаторам публикаций (Accession Number Web of Science) были выгружены данные, включающие следующий набор полей:

- Article Title (название публикации);
- Abstract (аннотация публикации);
- Author Keywords (авторские ключевые слова – присваиваются авторами публикаций);
- Keywords Plus (ключевые слова, присваиваемые автоматически в WoS: выделяются из заголовков публикаций, на которые ссылаются авторы статьи, каждый термин должен встречаться в библиографии более одного раза).

На основе Author Keywords и Keywords Plus был построен граф с помощью VOSviewer. Для этого сформирована выборка из 1123 наиболее часто встречающихся ключевых слов (каждое из них встречалось минимум 5 раз). Использовался фракционный счет, то есть вес каждого ключевого слова зависел от количества ключевых слов в каждой публикации. Например, если в публикации 10 ключевых слов, то вес у каждого из них равен 1/10. Для группировки (слияния) схожих терминов был использован тезаурус, предоставленный разработчиками VOSviewer. Масштаб (Scale) равен 0,9. Размер шрифта слов на *рисунке 1* характеризует частоту их встречаемости в выборке: чем больше размер шрифта, тем чаще встречается слово (для наглядности применен вариант

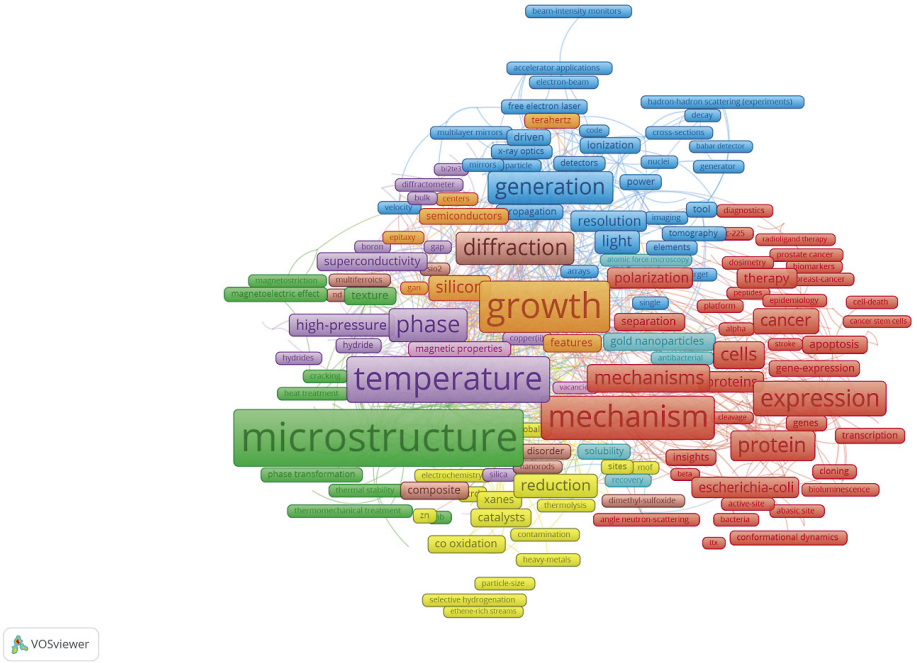

Рисунок 1. **Распределение Author Keywords и Keywords Plus, извлеченных из публикаций, авторами которых являются участники конкурса в рамках Программы, в зависимости от частоты встречаемости**

Источник: *Web of Science Core Collection, визуализация с помощью VOSviewer, данные на 28.12.2021 г.*

с наибольшей из возможных вариативностью размера шрифта, отображается в VOSviewer как 1.0). По остальным параметрам были выбраны значения по умолчанию. Расстояние между отдельными словами определяет степень тематических связей, в том числе их нахождение в одних и тех же публикациях: чем ближе слова расположены друг к другу, тем больше они тематически связаны между собой. Кроме того, по такому же принципу объединены слова в девять кластеров, обозначенных отдельными цветами, в каждом из которых наблюдается наибольшая концентрация связей между ключевыми словами. Обращает на себя внимание высокая частота встречаемости ключевых слов, не характеризующих специфику области синхротронных и нейтронных исследований и тем, предложенных в Программе: microstructure, growth, temperature, mechanism, expression и др. *(рисунок 1)*.

Для Author Keywords и Keywords Plus, извлеченных из публикаций, авторами которых являются победители конкурса в рамках Программы, сформирована выборка из 492 наиболее часто встречающихся ключевых слов по тем же принципам. В данном случае алгоритмы VOSviewer определили десять тематических кластеров, обозначенных отдельными цветами. В них чаще всего также встречаются ключевые слова, не характеризующие специфику области синхротронных и нейтронных исследований и тем Программы: spectroscopy, nanoparticles, films, temperature, expression *(рисунок 2)*.

Ограничение анализа тематической структуры публикаций только обработкой ключевых слов не обеспечивает глубокую смысловую интерпретацию результатов [26]. Детальный анализ тематических кластеров и взаимосвязей между различными ключевыми словами для отбора наиболее релевантных тем, как правило, требует участия экспертов [27]. Выявление наиболее важных групп связей между ключевыми словами, характеризующими синхротронные и нейтронные исследования, позволило бы рассмотреть полученные

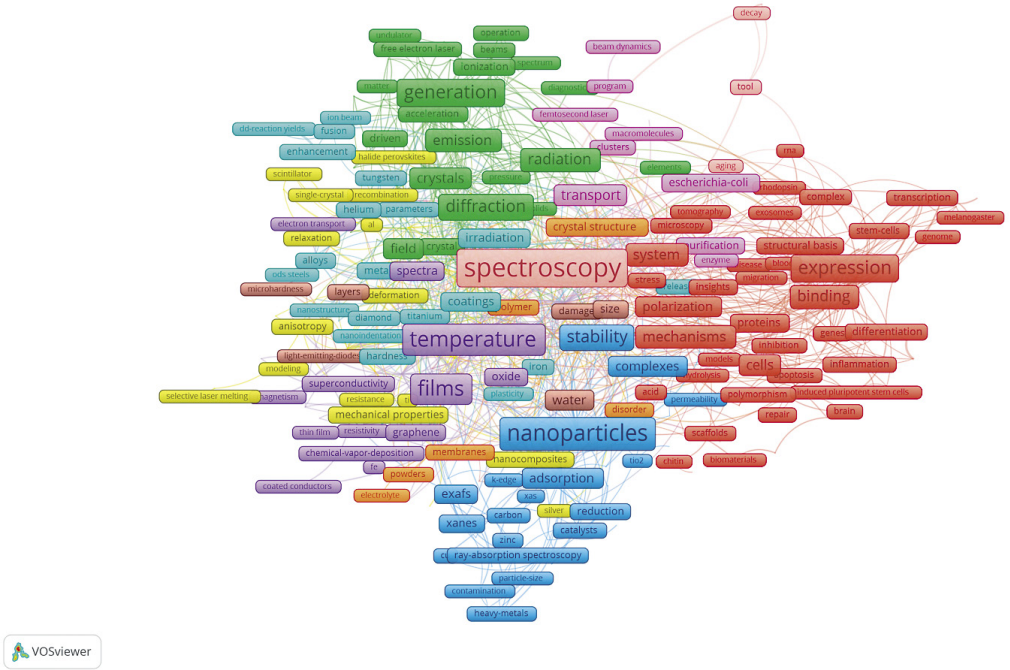

Рисунок 2. **Распределение Author Keywords и Keywords Plus, извлеченных из публикаций, авторами которых являются победители конкурса в рамках Программы, в зависимости от частоты встречаемости**

Источник: *Web of Science Core Collection, визуализация с помощью VOSviewer, данные на 28.12.2021 г.*

результаты подробнее, например, с точки зрения активно развивающихся тематик и направлений исследований.

В связи с тем, что предложенный поисковый запрос охватывает поля Article Title и Abstract, для отражения их семантических особенностей с помощью метода YAKE [21] были выделены ключевые слова из названий публикаций и их аннотаций. Для более релевантного отражения тенденций в научных публикациях рекомендуется использовать выделение ключевых слов из полных текстов [24]. Но в данной работе важно идентифицировать те ключевые слова, которые вероятнее всего будут содержаться в тех же полях, что используются для оценки публикационной результативности Программы. Именно поэтому часто недоступные полные тексты публикаций в данном случае не использовались. Полученные таким образом ключевые слова совместно с Author Keywords и Keywords Plus были объединены в общий список. Итого получено 95505 ключевых слов, которые встречаются в общей сложности 184124 раза.

Анализ частоты встречаемости терминов в результирующей выборке свидетельствует, что ключевые слова поискового запроса в WoS, предложенного экспертами, представлены незначительно, за исключением термина synchrotron, а именно:

- spallat* neutron source – ни разу не встречается;
  - spallat* – встречается 9 раз (5 раз как отдельное ключевое слово, остальные – в составе других ключевых слов);
  - neutron source – 25 раз (7 раз как отдельное ключевое слово, остальные – в составе других ключевых слов);
- neutron scattering – 137 раз (58 раз как отдельное ключевое слово, остальные – в составе других ключевых слов);
- neutron diffraction – 75 раз (51 раз как отдельное ключевое слово, остальные – в составе других ключевых слов);
- synchrotron – 584 раза (60 раз как отдельное ключевое слово, остальные – в составе других ключевых слов).

Если же обратиться к публикациям, авторами которых являются победители конкурса, то в них удалось идентифицировать 59422 ключевых слова, которые встречаются 88629 раз. При этом ключевые слова из того же самого поискового запроса встречаются среди них еще реже:

- spallat* neutron source – ни разу не встречается;
  - spallat* – встречается 8 раз (5 раз как отдельное ключевое слово, остальные – в составе других ключевых слов);
  - neutron source – 18 раз (6 раз как отдельное ключевое слово, остальные – в составе других ключевых слов);
- neutron scattering – 68 раз (26 раз как отдельное ключевое слово, остальные – в составе других ключевых слов);
- neutron diffraction – 38 раз (27 раз как отдельное ключевое слово, остальные – в составе других ключевых слов);
- synchrotron – 336 раз (39 раз как отдельное ключевое слово, остальные – в составе других ключевых слов).

Дополнительно проанализировано распределение перечисленных ключевых слов по годам публикаций в Scopus. Поиск производился по полям TITLE-ABS-KEY, со следующими ограничениями: PUBYEAR > 2015, DOCTYPE «ar», благодаря чему найдено следующее количество публикаций:

- spallat* neutron source – 761;
- neutron scattering – 9928;
- neutron diffraction – 7452;
- synchrotron – 22582.

Анализируемые наборы публикаций, за исключением neutron scattering, характеризуются нестабильным ростом на временной шкале[1] *(рисунок 3)*. В свою очередь, для темы neutron scattering наблюдается стабильное развитие и рост интереса мирового научного сообщества.

По ядру РИНЦ был произведен аналогичный поиск на русском и английском языках (с 2016 г., в названии публикации, в аннотации и в ключевых словах, тип публикации – статьи в журналах, без учета морфологии),

[1] 2022 г. не отображен на изображениях из-за неполноты данных на дату анализа.

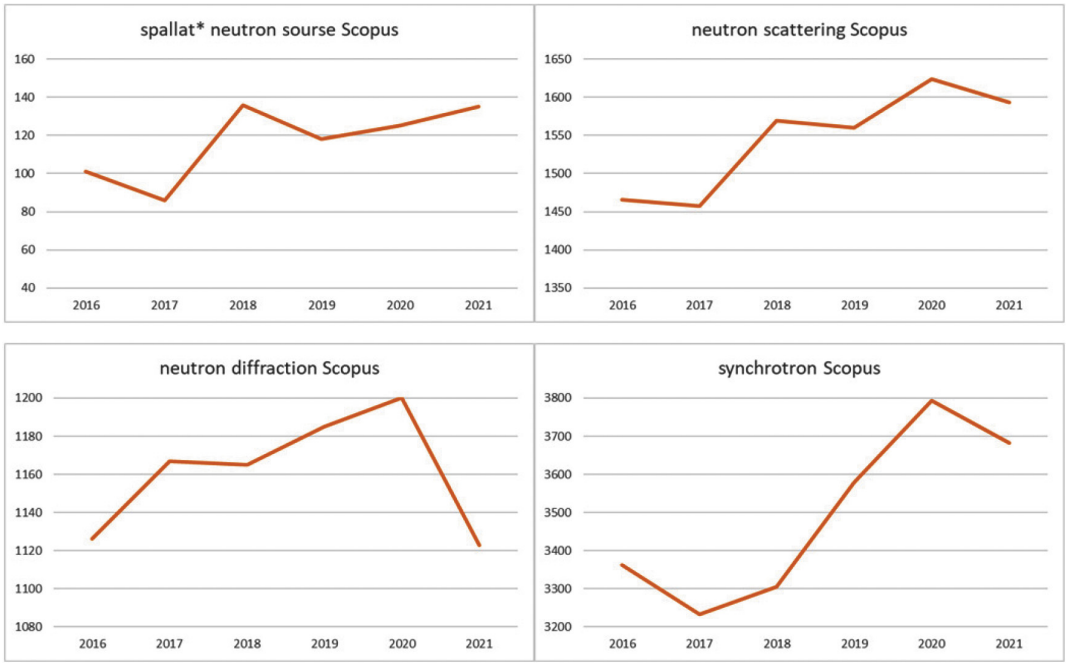

Рисунок 3. **Распределение по годам публикаций, содержащих ключевые слова из поискового запроса в Scopus, который составлен согласно Программе**
Источник: *Scopus, данные на 16.06.2022 г.*

благодаря чему идентифицировано следующее количество публикаций:

- spallat* neutron source, источник нейтронов делен* – 337;
- neutron scattering, рассеяние нейтронов – 2609;
- neutron diffraction, дифракция нейтронов – 2047;
- synchrotron, синхротрон – 4735.

В ядре РИНЦ на временной шкале (spallat* neutron source, источник нейтронов делен*) и (neutron diffraction, дифракция нейтронов) характеризуются стагнацией, а (neutron scattering, рассеяние нейтронов) и (synchrotron, синхротрон) – стабильным ростом, что говорит о развитии данных тем, но в то же время и об отсутствии среди них прорывных технологий *(рисунок 4)*.

Таким образом, результаты анализа позволяют с большой долей вероятности предположить, что поисковые запросы как в WoS, так и в Scopus, и ядре РИНЦ, необходимо расширить, используя дополнительные ключевые слова, отражающие развитие прорывных технологий для наиболее полного охвата публикаций в исследуемой области.

В связи с тем, что высокая частота встречаемости ключевых слов не всегда может характеризовать специфику анализируемой области (зашумление выборки за счет общеупотребимых терминов), к работе со списком ключевых слов должны быть привлечены эксперты в соответствующей области. В текущей работе в качестве примера для отработки алгоритма были отобраны термины, упоминаемые при описании научных направлений и ожидаемых результатов Программы, связанные с применением ускорителей протонов, ионов, электронов, а именно:

- «технологии ускорителей электронов, необходимые для создания новых источников синхротронного излучения 3-го и 4-го поколений...»;
- «технологии ускорителей протонов и ионов, необходимые для создания нейтронных источников...» и др [14].

Данные технологии характеризуются прежде всего термином accelerat* (то есть, acceleration, accelerator, accelerating, accelerated и др.), который в результирующей выборке встречается 56 раз как отдельное ключевое слово и 194 раза в составе 59422 других ключевых слов, в том числе:

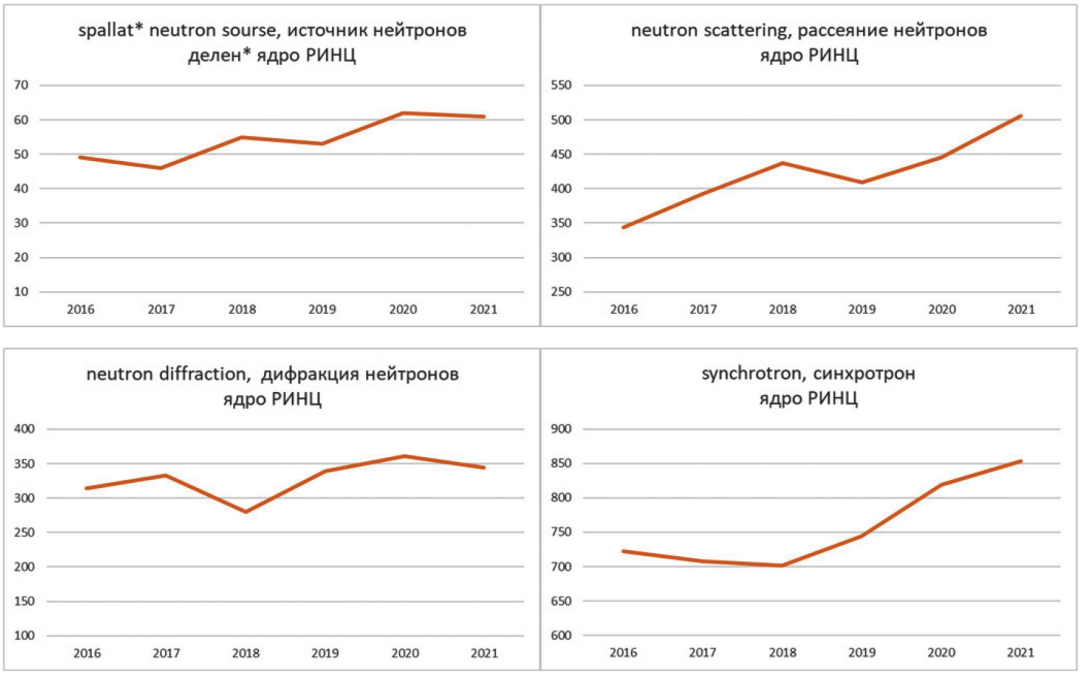

Рисунок 4. **Распределение по годам публикаций, содержащих ключевые слова из поискового запроса в ядре РИНЦ, который составлен согласно Программе**
Источник: *Ядро РИНЦ, данные на 16.06.2022 г.*

- ion accelerat* – 11 раз как отдельное ключевое слово и 16 раз в составе других ключевых слов;
- electron accelerat* 9 раз как отдельное ключевое слово и 13 раз в составе других ключевых слов;
- proton accelerat* 8 раз как отдельное ключевое слово и 5 раз в составе других ключевых слов.

Таким образом, ключевые слова, характеризующие ускорительные технологии, при условии дополнительной оценки уровня заинтересованности научного сообщества в их развитии, могут служить основой для уточнения поискового запроса. Для определения синтаксических рамок анализируемой научно-технологической области в поисковом запросе также проведен подбор операторов близости и усечений в Scopus и ядре РИНЦ.

При проведении поиска в Scopus с учетом новых терминов (по полям TITLE-ABS-KEY, с ограничениями: PUBYEAR > 2015, DOCTYPE «ar», W/2 – расстояние между словами ключевыми составляет максимум 2 слова) найдено следующее количество публикаций:

- ion accelerat* (ion W/2 accelerat*) – 2960;
- electron accelerat* (electron W/2 accelerat*) – 5633;
- proton accelerat* (proton W/2 accelerat*) – 1403.

Распределение публикаций, содержащих ion accelerat* (ion W/2 accelerat*), electron accelerat* (electron W/2 accelerat*), по годам характеризуется стабильным ростом, а proton accelerat* (proton W/2 accelerat*) практически стагнацией, что говорит об ускорителях ионов и электронов как активно развивающихся темах исследований на мировом уровне *(рисунок 5)*.

По ядру РИНЦ удалось провести двуязычный поиск только с более жесткими ограничениями в виде точных фраз, заключенных в кавычки (с 2016 г., поиск в названии публикации, аннотации, ключевых словах; тип публикации – статьи в журналах, без учета морфологии). Возможность более гибкого расположения отдельных терминов на расстоянии нескольких слов друг от друга на платформе eLIBRARY.RU не предусмотрена.

При таких условиях выявлено следующее количество публикаций:

- ion accelerat* (“ion accelerat*” OR “ускор* ион*”) – 1001;

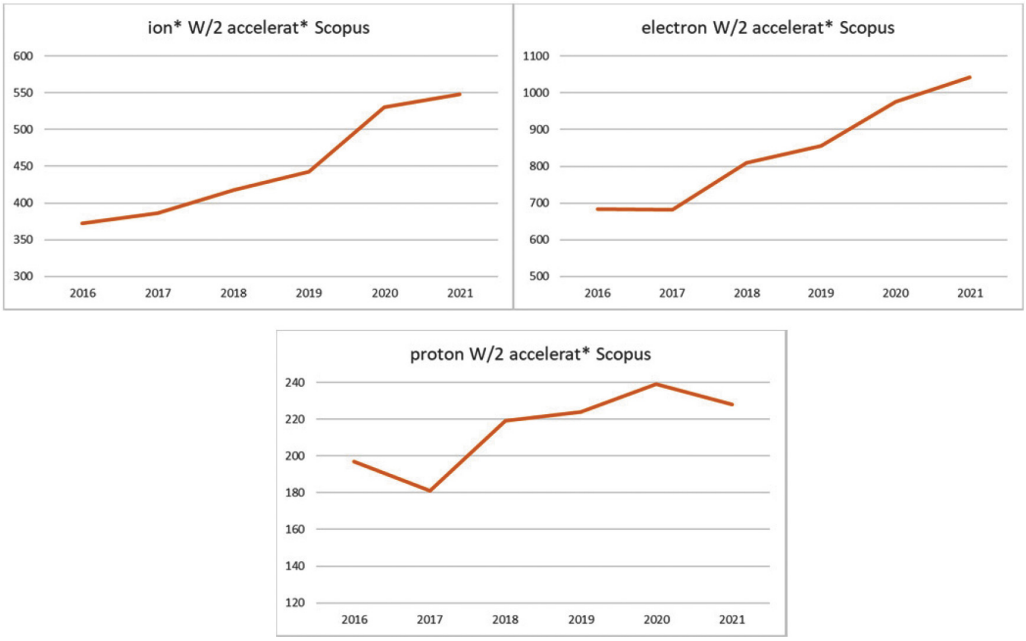

Рисунок 5. **Распределение по годам публикаций, содержащих новые ключевые слова для уточненного поискового запроса в Scopus**
Источник: *Scopus, данные на 16.06.2022 г.*

- electron accelerat* ("electron accelerat*" OR "ускор* электрон*") – 1611;
- proton accelerat* ("proton accelerat*" OR "ускор* протон*") – 533.

Все три анализируемые темы в ядре РИНЦ характеризуются стагнацией *(рисунок 6).*

Для оценки влияния синтаксиса поискового запроса на характер формируемых выборок

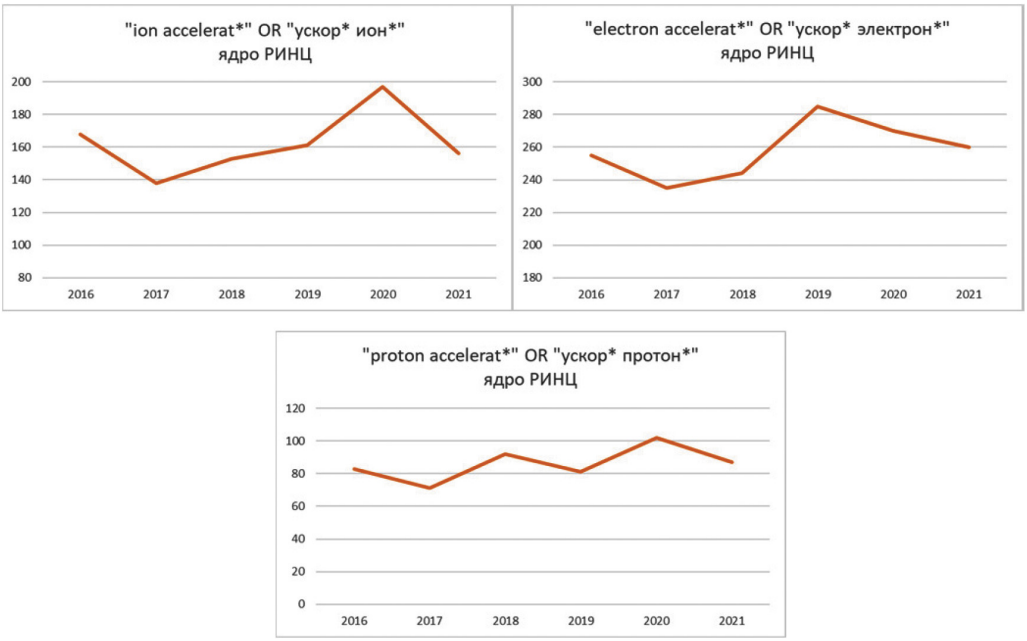

Рисунок 6. **Распределение по годам публикаций, содержащих новые ключевые слова для уточненного поискового запроса в ядре РИНЦ**
Источник: *Ядро РИНЦ, данные на 16.06.2022 г.*

был проведен поиск с идентичными ограничениями (по точным фразам) в Scopus. В результате найдены публикации, распределение которых на временной шкале также, как и в ядре РИНЦ, характеризуется стагнацией *(рисунок 7)*:

- ion accelerat* («ion accelerat*») – 1210;
- electron accelerat* («electron accelerat*») – 1657;
- proton accelerat* («proton accelerat*») – 684.

При этом использование точных фраз не охватывает все представленные публикации в анализируемых областях, поэтому применение ядра РИНЦ в качестве базы данных для поиска позволит получать более полные и релевантные коллекции лишь при появлении возможности гибкой настройки близости расположения отдельных терминов.

Благодаря добавлению новых ключевых слов, характеризующих технологии ускорителей ионов, электронов и протонов, применению операторов усечения и логических операторов, обозначающих близость расположения ключевых слов друг от друга, был сформирован следующий поисковый запрос в WoS, отражающий рамки анализируемой научно-технологической области:

TS=((spallat* AND (neutron* NEAR/2 source*)) OR (neutron* NEAR/2 scatter*) OR (neutron* NEAR/2 diffract*) OR (synchrotron*) OR (accelerat* NEAR/2 (ion* OR electron* OR proton*))) Refined by: DOCUMENT TYPES: (ARTICLE) AND COUNTRIES/REGIONS: (RUSSIA)

В Scopus аналогичный поисковый запрос выглядит следующим образом:

TITLE-ABS-KEY ((spallat* AND (neutron* W/2 source*)) OR (neutron* W/2 scatter*) OR (neutron* W/2 diffract*) OR (synchrotron*) OR (accelerat* W/2 (ion* OR electron* OR proton*))) AND (LIMIT-TO (AFFILCOUNTRY, "Russian Federation")) AND (LIMIT-TO (DOCTYPE, "ar"))

В ядре РИНЦ с учетом всех существующих ограничений возможно составить такой поисковый запрос:

«spallat* neutron* source*» OR «источник* нейтрон* делен*» OR «neutron* scatter*» OR «рассеян* нейтрон*» OR «neutron* diffract*» OR «дифракц* нейтрон*» OR «synchrotron*» OR «синхротрон*» OR «ion* accelerat*» OR «ускор* ион*» OR «electron* accelerat*» OR «ускор*

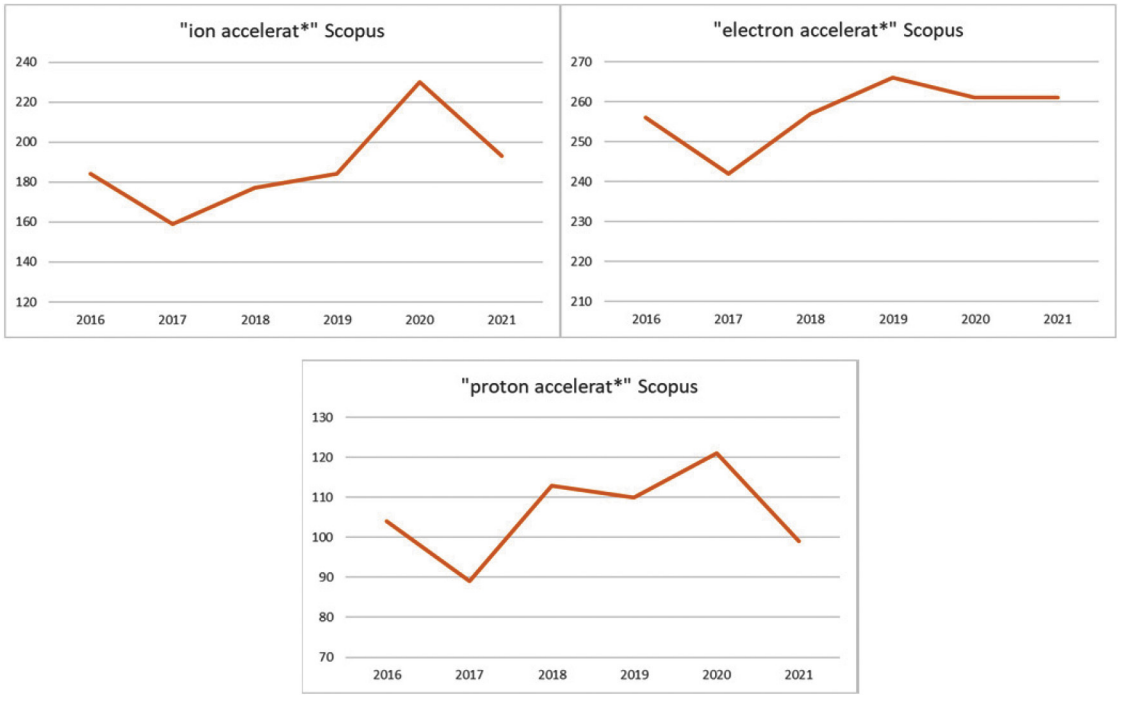

Рисунок 7. **Распределение по годам публикаций, содержащих новые ключевые слова в виде точных фраз для уточненного поискового запроса в Scopus**

Источник: *Scopus, данные на 16.06.2022 г.*

электрон*» OR «proton* accelerat*» OR «ускор* протон*» (поиск в названии публикации, в аннотации, в ключевых словах, тип публикации – статьи в журналах, без учета морфологии)

Результаты поиска по указанным трем вариантам уточненных запросов представлены в столбцах 6–8 *таблицы 1*.

В *таблице 1* приведены плановые и выявленные значения по индикатору И1 в зависимости от вида поискового запроса и выбранной базы данных в качестве источника анализируемой информации.

Проведено сравнение количества найденных публикаций за первые три года реализации Программы (2019–2021 гг.) с плановыми значениями И1. Наименьшее отставание найденных значений И1 от плановых наблюдается в 2021 г. по уточненному поисковому запросу в WoS с применением дополнительных ключевых слов (619 при плановом значении 630; отставание – 11 публикаций). Причем в 2019 и 2020 гг. по этому же поисковому запросу выявлено превышение планового индикатора на 70 и 95 публикаций, соответственно. На втором месте, с отставанием на 74 публикации в 2021 г. – уточненный поисковый запрос в Scopus. На третьем и четвертом месте оказались запросы в WoS и Scopus согласно предложенным в Программе темам: наибольшее запаздывание числа найденных публикаций от планового значения отмечено по ним в 2021 г. – соответственно, на 107 и 180 единиц. Поисковый запрос в ядре РИНЦ, соответствующий предложенному в Программе, позволил найти количество публикаций, значительно превышающее плановые значения И1 (в среднем, на 125 публикаций). Уточненный поисковый запрос выявил еще большее количество публикаций, в частности, в 2020 г. – более 1000 при плановом значении 600 публикаций.

Уточненные поисковые запросы в WoS и Scopus позволяют получить данные, более релевантные плановым значениям И1 (с меньшим отставанием) по сравнению с поисковыми запросами, представленными в рамках Программы. Поисковые запросы в ядре РИНЦ характеризуются большим количеством публикаций,

Таблица 1

Значения индикатора «Количество публикаций в области синхротронных и нейтронных исследований (разработок) в журналах, индексированных в международных базах данных» в зависимости от вида поискового запроса и выбора базы данных, единиц

Год	План	Факт					
		Поисковый запрос согласно Программе			Уточненный поисковый запрос		
		WoS, 29.04.2022 г.	Scopus, 10.06.2022 г.	Ядро РИНЦ, 14.06.2022 г.	WoS, 29.04.2022 г.	Scopus, 16.06.2022 г.	Ядро РИНЦ, 16.06.2022 г.
1	2	3	4	5	6	7	8
2027	960	-	-	-	-	-	-
2026	860	-	-	-	-	-	-
2025	800	-	-	-	-	-	-
2024	750	-	-	-	-	-	-
2023	710	-	-	-	-	-	-
2022	670	108	167	160	141	220	202
2021	630	523	450	768	619	556	985
2020	600	557	510	780	695	673	1023
2019	570	534	465	747	640	614	922
2018	-	550	441	513	668	591	686
2017	-	510	423	523	606	538	669
2016	-	510	422	521	622	551	638

Источник: *Web of Science Core Collection, Scopus, ядро РИНЦ, данные на указанную в столбцах дату*

чем в WoS и Scopus, что объясняется главным образом спецификой ядра РИНЦ – охватом более высокого числа российских научных журналов (RSCI) и ограниченными функциями настройки по синтаксису поисковых запросов на платформе eLIBRARY.RU. В целом использование ядра РИНЦ для оценки публикационной результативности Программы возможно после изменения либо плановых значений по количеству публикаций, либо ряда поисковых функций на платформе eLIBRARY.RU.

ЗАКЛЮЧЕНИЕ

Практика ФГБНУ «Дирекция НТП» в проведении исследований публикационной результативности в рамках различных федеральных программ показывает, что получение максимально качественной коллекции публикаций по определенной тематике для ее последующего анализа – сложный и многоэтапный процесс, в ходе которого не вполне точно составленный поисковый запрос может привести как к техническим ошибкам поиска, так и к получению нерелевантной и неполной выборки. Анализ вариантов терминологического и синтаксического наполнения поискового запроса, необходимого для оценки публикационной результативности в области синхротронных и нейтронных исследований и разработок, показал, что публикации российских ученых, участвующих в конкурсе Программы, содержат обширный массив ключевых слов для анализа и построения релевантного тематического поискового запроса и решения задачи наиболее полного охвата пула публикаций в исследуемой научно-технологической области. Как и предполагалось, подготовленная результирующая выборка ключевых слов из публикаций участников и победителей конкурса была значима для валидации представленного в Программе поискового запроса.

Исследование показало, что примерно половина тем Программы, связанных с «импульсными источниками нейтронов на основе реакции испарительно-скалывающего типа» и «дифракцией нейтронов», характеризуются нестабильным ростом или даже стагнацией (по данным из ядра РИНЦ). А для тем «рассеяние нейтронов» и «синхротрон» наблюдается стабильный рост интереса мирового научного сообщества, что говорит об их развитии. С использованием дополнительных терминов, например, описывающих технологии ускорителей ионов, электронов и протонов, можно осуществить поисковый запрос с выдачей более полного результата, отражающего иные развивающиеся тематики в области применения синхротронного и нейтронного излучения.

В статье предложен и апробирован алгоритм выделения и уточнения рамок научно-технологической области в различных библиометрических базах данных на примере доработки поискового запроса Программы для более полной и точной оценки ее публикационной результативности. Основными этапами алгоритма являются:

1) однозначная идентификация и верификация всех публикаций из заявок в библиометрических базах данных с помощью предварительной обработки с применением API;

2) оценка возможности воспроизведения предложенного в рамках Программы поискового запроса в различных библиометрических базах данных;

3) составление списка ключевых слов, содержащихся в публикациях участников и победителей конкурса в рамках Программы (из всех полей, предусмотренных в анализируемом поисковом запросе):

- выгрузка Article Title (название публикации), Abstract (аннотация публикации), Author Keywords и Keywords Plus (при наличии);
- построение графов в VOSviewer по Author Keywords и Keywords Plus (при наличии) отдельно для участников и для победителей конкурса, выявление и оценка наиболее часто встречающихся ключевых слов;
- выделение ключевых слов из Article Title, Abstract с помощью метода текст-майнинга YAKE;
- формирование единого списка ключевых слов;

4) подсчет частоты встречаемости терминов из поискового запроса, предложенного в рамках Программы, в полученном списке ключевых слов;

5) оценка частоты встречаемости терминов из поискового запроса, предложенного в рамках

Программы, на временной шкале при поиске в различных библиометрических базах данных;

6) выбор новых ключевых слов по релевантным Программе научно-технологическим областям;

7) анализ новых ключевых слов, аналогичный проводимому на этапах 4–5;

8) определение рамок научно-технологической области на основе уточненных поисковых запросов в различных библиометрических базах с использованием новых ключевых слов и особенностей синтаксиса в каждой из них;

9) выбор наиболее подходящего поискового запроса.

Описанный алгоритм в дальнейшем может использоваться при выделении ключевых слов на этапе 6 по отдельным научно-технологическим областям Программы (например, реакторные технологии) в совокупности с экспертной оценкой. Подобный подход применим не только в отношении сформированного в текущей работе списка ключевых слов, но и в списках, которые можно создавать на основании данных из публикаций, представляемых в новых заявках и отчетных материалах исполнителей проектов. Выделенные с помощью экспертов ключевые слова могут быть использованы для составления новых уточненных поисковых запросов не только в международных библиометрических базах данных (доступ к которым в России либо уже приостановлен, либо может быть прекращен), но и в доступных отечественных информационных ресурсах, например, в ядре РИНЦ на платформе eLIBRARY.RU. Привлечение экспертов на других этапах позволит проводить валидацию поискового запроса с более глубоким пониманием специфики анализируемых научно-технологических областей. Выполнение подобной работы планируется при проведении дальнейших исследований в ходе организационно-методического сопровождения реализации Программы.

Проводимая на этапе 1 описанного алгоритма предварительная обработка публикационных данных с применением API WoS, Scopus также может быть использована для более оперативного анализа публикаций, созданных в ходе реализации Программы. Модификация этой обработки применима при работе и с другими информационными ресурсами, например, eLIBRARY.RU.

Весной 2022 г. компания CrossRef перестала регистрировать новые префиксы и новые коды DOI для части российских организаций в условиях санкций [28]. То есть часть публикаций из российских журналов смогут содержать DOI, а часть – нет. Наличие DOI или других аналогичных идентификаторов является важным условием для обработки публикационных данных в полуавтоматическом режиме. В случае работы с платформой eLIBRARY.RU проблема отсутствия DOI частично решаема. Для этого в п. 1.4 приложения 2 к форме 4 «Сведения о квалификации» конкурсной документации и других похожих формах Программы следует добавить eLIBRARY Document Number (EDN) – уникальный код, который присваивается всем документам в данном информационном ресурсе. EDN позволит провести идентификацию и верификацию всех публикаций, вышедших в журналах из ядра РИНЦ или на любом другом уровне агрегации данных, представленном на платформе eLIBRARY.RU. Но российским журналам важно продолжать присваивать публикациям DOI через Crossref при наличии такой возможности. В отличие от EDN, DOI увеличивает видимость российских публикаций для читателей во всем мире и позволяет включать публикации не только в отечественные, но и в зарубежные библиометрические исследования.

Научная электронная библиотека (НЭБ) и Российская академия наук (РАН) 26.05.2022 г. подписали соглашение о сотрудничестве, в рамках которого планируется активная работа по развитию ядра РИНЦ. В том числе предполагается «разработка и развитие методологического аппарата, отдельных методик и инструментальных средств для экспертизы и оценки научных достижений на основе соединения формальных статистических методов науко- и библиометрических измерений и качественных методов экспертного анализа» [29]. Информация из настоящей работы может быть использована для улучшения поискового функционала платформы eLIBRARY.RU, а также послужить основой для решения упомянутых задач, которые стоят не только перед РАН и НЭБ, но и всеми акторами современной научной политики России.

ЛИТЕРАТУРА

1. *Дежина И.Г.* (2020) Трансформационные исследования: новый приоритет государств после пандемии. – М.: Издательство Ин-та Гайдара. 116 с.
2. *Гуськов А.Е., Косяков Д.В.* (2020) Национальный фракционный счет и оценка научной результативности организаций // Научные и технические библиотеки. 1(9):15–42. DOI:10.33186/1027-3689-2020-9-15-42.
3. *Стерлигов И.А.* (2021) Российский конференционный взрыв: масштабы, причины, дальнейшие действия // Управление наукой: теория и практика. 3(2):222–251. DOI: 10.19181/smtp.2021.3.2.10.
4. Руководство по наукометрии: индикаторы развития науки и технологии, второе издание (2021) / М.А. Акоев, В.А. Маркусова, О.В. Москалева, В.В. Писляков; под. ред. М.А. Акоева; Екатеринбург: Издательство Уральского университета. 358 с. DOI: 10.15826/B978-5-7996-3154-3.
5. *Жэнгра И.* (2018) Ошибки в оценке науки, или Как правильно использовать библиометрию; пер. с франц. А. Зайцевой. – М.: Новое литературное обозрение. 184 с.
6. *Шомшор М., Адамс Д., Пендлбери Д.А., Роджерс Г.* (2021) Классификация данных: как делать осознанный выбор, ведущий к желаемым результатам / Отчет о международном исследовании Института научной информации. – https://discover.clarivate.com/data_categorization_ru?utm_campaign=EM1_ISI_11_GRR_InCites_Data_Categorization_LeadGen_SAR_RussiaCIS_2021&utm_medium=email&utm_source=Eloqua.
7. *Abramo G., D'Angelo C.A., Zhang L.* (2018) A comparison of two approaches for measuring interdisciplinary research output: The disciplinary diversity of authors vs the disciplinary diversity of the reference list // Journal of Informetrics. 12(4):1182–1193. DOI: 10.1016/j.joi.2018.09.001.
8. *Zhang L., Sun B., Jiang L., Huang Y.* (2021) On the relationship between interdisciplinarity and impact: Distinct effects on academic and broader impact // Research Evaluation. 30(3):256–268. DOI: 10.1093/reseval/rvab007.
9. *Pech G., Delgado C., Sorella S.P.* (2022) Classifying papers into subfields using Abstracts, Titles, Keywords and KeyWords Plus through pattern detection and optimization procedures: An application in Physics // Journal of the Association for Information Science and Technology. Article in Press:1–16. DOI: 10.1002/asi.24655.
10. *Thijs B., Zhang L., Glänzel W.* (2015) Bibliographic coupling and hierarchical clustering for the validation and improvement of subject-classification schemes // Scientometrics. 105(3):1453–1467. DOI: 10.1007/s11192-015-1641-3.
11. *Bode C., Herzog C., Hook D., McGrath R.* (2018) A guide to the dimensions data approach. A collaborative approach to creating a modern infrastructure for data describing research: where we are and where we want to take it. London: Digital Science. DOI: 10.6084/m9.figshare.5783094.v7.
12. *Bornmann L.* (2018). Field classification of publications in dimensions: A first case study testing its reliability and validity // Scientometrics. 117(1):637–640. DOI: 10.1007/s11192–018–2855-y.
13. *Herzog C., Lunn B.K.* (2018) Response to the letter "Field classification of publications in dimensions: A first case study testing its reliability and validity" // Scientometrics. 117(1):641–645. DOI:10.1007/s11192–018–2854-z.
14. Постановление Правительства Российской Федерации от 16.03.2020 г. № 287 (2020) Федеральная научно-техническая программа развития синхротронных и нейтронных исследований и исследовательской инфраструктуры на 2019–2027 годы. http://publication.pravo.gov.ru/Document/View/0001202003260022
15. *Благов А.Е.* (2021) Источники синхротронного излучения четвертого поколения и лазеры на свободных электронах – основа современной кристаллографии и материаловедения / Заседание Президиума РАН 14.09.2021 г. «Научная Россия», 14.09.2021. https://scientificrussia.ru/articles/zasedanie-prezidiuma-ran-14092021.
16. Путин утвердил сроки создания синхротронных и нейтронных мегаустановок (2019) / РИА, 25.07.2019. https://ria.ru/20190725/1556871808.html.
17. Чернышенко одобрил проект создания научной установки на острове Русский (2021) / РИА, 10.12.2021. https://ria.ru/20211210/megasayens-1763194751.html.
18. Мегазапуск: утвержден план развития синхротронных исследований (2019) / Известия, 23.10.2019. https://iz.ru/935014/dmitrii-istomin/megazapusk-utverzhden-plan-razvitiia-sinkhrotronnykh-issledovanii.
19. Постановление Правительства Российской Федерации от 19.03.2022 г. № 414 (2022) О некоторых вопросах применения правовых актов Правительства Российской Федерации, устанавливающих требования, целевые значения показателей по публикационной активности / Официальный интернет-портал правовой информации. http://publication.pravo.gov.ru/Document/View/0001202203210040.
20. Конкурсная документация по проведению конкурса на предоставление грантов в форме субсидий из федерального бюджета на реализацию отдельных мероприятий Федеральной научно-

технической программы развития синхротронных и нейтронных исследований и исследовательской инфраструктуры на 2019–2027 годы (2021) Утв. Заместителем Министра науки и высшего образования Российской Федерации А.М. Медведевым 20 мая 2021 г.

21. *Campos R., Mangaravite V., Pasquali A., Jatowt A., Jorge A., Nunes C., Jatowt A.* (2020) YAKE! Keyword Extraction from Single Documents using Multiple Local Features // Information Sciences Journal. 509:257–289. DOI: 10.1016/j.ins.2019.09.013.

22. *Borner K., Chen C.M., Boyack K.W.* (2003) Visualizing knowledge domains // Annual Review of Information Science and Technology. 37:179–255. DOI: 10.1002/aris.1440370106.

23. *Mohammadi E., Karami A.* (2022) Exploring research trends in big data across disciplines: A text mining analysis // Journal of Information Science. 48(1):44–56. DOI: 10.1177/0165551520932855.

24. *Rezaeian M., Montazeri H., Loonen R.C.G.M.* (2017) Science foresight using life-cycle analysis, text mining and clustering: A case study on natural ventilation // Technological Forecasting and Social Change. 118:270–280. DOI: 10.1016/j.techfore.2017.02.027.

25. Fundamentals of predictive text mining (2010) / S.M. Weiss, N. Indurkhya, T. Zhang; Springer. 226 p. DOI: 10.1007/978-1-84996-226-1.

26. *Daim T., Bukhari E., Bakry D., VanHuis J., Yalcin H., Wang X.* (2021) Forecasting Technology Trends through the Gap Between Science and Technology: The Case of Software as an E-Commerce Service // Foresight and STI Governance. 2021; 15(2):12–24. DOI: 10.17323/2500–2597.2021.2.12.24.

27. *Солошенко Н.С., Пронина Т.А, Зибарева И.В.* (2017) Возможности использования лингвистических аппаратов реферативно-аналитических ресурсов при выявлении новых направлений в междисциплинарных научных исследованиях: библиометрический подход / Информация в современном мире. Международная конференция, посвящается 65-летию ВИНИТИ РАН. Материалы конференции. Москва, 2017. 286–297 с.

28. Вице-президент РАН Алексей Хохлов: DOI-импортозамещение (2022) / Поиск, 06.04.2022. https://poisknews.ru/science-politic/vicze-prezident-ran-aleksej-hohlov-doi-importozameshhenie.

29. Пресс-релиз о подписании соглашения о сотрудничестве между РАН и НЭБ (2022) / Научная электронная библиотека eLIBRARY.RU, 26.05.2022. https://elibrary.ru/projects/rsci/ran_2022.pdf.

Информация об авторах

Чернова Ирина Николаевна – кандидат исторических наук, старший научный сотрудник отдела аналитических исследований, ФГБНУ «Дирекция НТП» (Российская Федерация, 123557, г. Москва, ул. Пресненский вал, д. 19; e-mail: chernova@fcntp.ru).

Черченко Ольга Владимировна – научный сотрудник отдела аналитических исследований, ФГБНУ «Дирекция НТП»; Scopus Author ID: 57209975440, ORCID: 0000-0002-2669-0885 (Российская Федерация, 123557, г. Москва, ул. Пресненский вал, д. 19; e-mail: olya.cherchenko@mail.ru).

I.N. CHERNOVA,
SSTP Directorate (Moscow, Russian Federation; e-mail: chernova@fcntp.ru)
O.V. CHERCHENKO,
SSTP Directorate (Moscow, Russian Federation; e-mail: olya.cherchenko@mail.ru)

THE ALGORITHM FOR REFINING A FRAMEWORK OF SCIENTIFIC AND TECHNOLOGICAL FIELD IN BIBLIOMETRIC DATABASES ON THE EXAMPLE OF SYNCHROTRON, NEUTRON RESEARCH AND DEVELOPMENT

UDC: 001
10.22394/2410-132X-2022-8-2-98-117

Abstract: To assess the publication effectiveness of the Federal Scientific and Technical Program for the Development of Synchrotron and Neutron Research and Research Infrastructure for 2019–2027 (Program), its developers proposed the special search query in Web of Science Core Collection (WoS). The purpose of the study is to determine effective approaches to adaptation of the search query providing the most complete coverage of the Program topics

publications and applicable to international and domestic databases. In publications from applications for participation in the Program, we identified keywords to validate the search query. We used keyword search in WoS, Scopus, core of Russian Index of Science Citation, identification of publications by using WoS and Scopus APIs, graph building in VOSviewer, Yet Another Keyword Extractor text mining method. On the basis of empirical data, a multistage algorithm was proposed to the formation of a specific scientific and technological field collection of publications in bibliometric databases.

Keywords: *indicators, publication activity, search query, bibliometric analysis, text-mining, synchrotron studies, neutron studies*

Acknowledgements: The work was supported financially by the Ministry of Education and Science of the Russian Federation under subsidy agreement No. 075-02-2022-979 dated February 22, 2022.

For citation: Chernova I.N., Cherchenko O.V. The Algorithm for Refining a Framework of Scientific and Technological Field in Bibliometric Databases on the Example of Synchrotron, Neutron Research and Development. *The Economics of Science*. 2022; 8(2):98–117. https://doi.org/10.22394/2410-132X-2022-8-2-98-117

REFERENCES

1. *Dezhina I.G.* (2020) Transformational research: new priority of the state after the pandemic. – M.: Publishing house of The Gaidar Institute. 116 p. (In Russ.)
2. *Guskov A.E., Kosyakov D.V.* (2020) National fractional calculations and evaluating organization's science efficiency // Scientific and Technical Libraries. 1(9):15–42. DOI: 10.33186/1027-3689-2020-9-15-42. (In Russ.)
3. *Sterligov I.A.* (2021) The Russian Conference Outbreak: Description, Causes and Possible Policy Measures // Science Management: Theory and Practice. 3(2): 222–251. DOI: 10.19181/smtp.2021.3.2.10. (In Russ.)
4. Handbook for Scientometrics: Indicators of science and technology development (2021) / M.A. Akoev, V.A. Markusova, O.V. Moskaleva, V.V. Pislyakov; под. ред. M.A. Akoev; Yekaterinburg: Publishing house of The Ural Federal University. 358 p. DOI: 10.15826/B978-5-7996-3154-3. (In Russ.)
5. *Gingras Y.* (2018) Bibliometrics and Research Evaluation: Uses and Abuses; transl. from French A. Zaitseva. – M.: The New Literary Review. 184 p. (In Russ.)
6. *Szomszor M., Adams J., Pendlebury D.A., Rogers G.* (2021) Data categorization: Understanding choices and outcomes / The Global Research Report from the Institute for Scientific Information. https://discover.clarivate.com/data_categorization_ru?utm_campaign=EM1_ISI_11_GRR_InCites_Data_Categorization_LeadGen_SAR_RussiaCIS_2021&utm_medium=email&utm_source=Eloqua. (In Russ.)
7. *Abramo G., D'Angelo C.A., Zhang L.* (2018) A comparison of two approaches for measuring interdisciplinary research output: The disciplinary diversity of authors vs the disciplinary diversity of the reference list // Journal of Informetrics. 12(4):1182–1193. DOI: 10.1016/j.joi.2018.09.001.
8. *Zhang L., Sun B., Jiang L., Huang Y.* (2021) On the relationship between interdisciplinarity and impact: Distinct effects on academic and broader impact // Research Evaluation. 2021; 30(3):256–268. DOI: 10.1093/reseval/rvab007.
9. *Pech G., Delgado C., Sorella S.P.* (2022) Classifying papers into subfields using Abstracts, Titles, Keywords and KeyWords Plus through pattern detection and optimization procedures: An application in Physics // Journal of the Association for Information Science and Technology. Article in Press: 1–16. DOI: 10.1002/asi.24655.
10. *Thijs B., Zhang L., Glänzel W.* (2015) Bibliographic coupling and hierarchical clustering for the validation and improvement of subject-classification schemes // Scientometrics. 105(3):1453–1467. DOI: 10.1007/s11192-015-1641-3.
11. *Bode C., Herzog C., Hook D., McGrath, R.* (2018) A guide to the dimensions data approach. A collaborative approach to creating a modern infrastructure for data describing research: where we are and where we want to take it. London: Digital Science. DOI: 10.6084/m9.figshare.5783094.v7.
12. *Bornmann L.* (2018). Field classification of publications in dimensions: A first case study testing its reliability and validity // Scientometrics. 117(1):637–640. DOI: 10.1007/s11192–018–2855-y.
13. *Herzog C., Lunn B.K.* (2018) Response to the letter "Field classification of publications in dimensions: A first case study testing its reliability and validity" // Scientometrics. 117(1):641–645. DOI: 10.1007/s11192–018–2854-z.
14. Decree of the Government of the Russian Federation dated 16.03.2020 № 287 (2020) The Federal Scientific and Technical Program for the Development of Synchrotron and Neutron Research and Research Infrastructure for 2017–2027. http://publication.pravo.gov.ru/Document/View/0001202003260022. (In Russ.)
15. *Blagov A.E.* (2021) Sources of synchrotron radiation of the fourth generation and free electron lasers – the basis of modern crystallography and materials science. Meeting of the Presidium of the Russian Academy of Sciences on 14.09.2021 / Scientific

Russia, 14.09.2021. https://scientificrussia.ru/articles/zasedanie-prezidiuma-ran-14092021. (In Russ.)
16. Putin approved the deadlines for the creation of synchrotron and neutron mega-installations (2019) / RIA, 25.07.2019. https://ria.ru/20190725/1556871808.html. (In Russ.)
17. Chernyshenko approved the project for the creation of a scientific installation on Russky Island (2021) / RIA, 10.12.2021. https://ria.ru/20211210/megasayens-1763194751.html. (In Russ.)
18. Megalaunch: Synchrotron research development plan approved (2019) / Izvestia, 23.10.2019. https://iz.ru/935014/dmitrii-istomin/megazapusk-utverzhden-plan-razvitiia-sinkhrotronnykh-issledovanii. (In Russ.)
19. Decree of the Government of the Russian Federation dated 19.03.2022 № 414 (2022) On some issues of application of legal acts of the Government of the Russian Federation, establishing requirements, target values of indicators for publication activity / Official Internet portal of legal information. http://publication.pravo.gov.ru/Document/View/0001202203210040. (In Russ.)
20. Competitive documentation for holding a competition for grants in the form of subsidies from the federal budget for the implementation of the Federal Scientific and Technical Program for the Development of Synchrotron and Neutron Research and Research Infrastructure for 2017–2027 (2021) Approved by the Deputy Minister of Science and Higher Education of the Russian Federation A.M. Medvedev May 20, 2021. (In Russ.)
21. *Campos R., Mangaravite V., Pasquali A., Jatowt A., Jorge A., Nunes C., Jatowt A.* (2020) YAKE! Keyword Extraction from Single Documents using Multiple Local Features // Information Sciences Journal. 2020; 509:257–289. DOI: 10.1016/j.ins.2019.09.013.
22. *Borner K., Chen C.M., Boyack K.W.* (2003) Visualizing knowledge domains // Annual Review of Information Science and Technology. 37:179–255. DOI: 10.1002/aris.1440370106.
23. *Mohammadi E., Karami A.* (2022) Exploring research trends in big data across disciplines: A text mining analysis // Journal of Information Science. 48(1):44–56. DOI: 10.1177/0165551520932855.
24. *Rezaeian M., Montazeri H., Loonen R.C.G.M.* (2017) Science foresight using life-cycle analysis, text mining and clustering: A case study on natural ventilation // Technological Forecasting and Social Change. 118:270–280. DOI: 10.1016/j.techfore.2017.02.027.
25. Fundamentals of predictive text mining (2010) / S.M. Weiss, N. Indurkhya, T. Zhang; Springer. 226 p. 10.1007/978-1-84996-226-1.
26. *Daim T., Bukhari E., Bakry D., VanHuis J., Yalcin H., Wang X.* (2021) Forecasting Technology Trends through the Gap Between Science and Technology: The Case of Software as an E-Commerce Service // Foresight and STI Governance. 2021; 15(2):12–24. DOI: 10.17323/2500–2597.2021.2.12.24.
27. *Soloshenko N.S., Pronina T.A., Zibareva I.V.* (2017) Possibilities of using the linguistic apparatus of abstract and analytical resources in identifying new directions in interdisciplinary scientific research: a bibliometric approach // Information in the modern world. International conference dedicated to the 65th anniversary of VINITI RAS. Conference materials. Moscow, 2017. 286–297 p.
28. RAS Vice President Alexei Khokhlov: DOI-import substitution (2022) / Search, 06.04.2022. https://poisknews.ru/science-politic/vicze-prezident-ran-aleksej-hohlov-doi-importozameshhenie.
29. Press release on the signing of a cooperation agreement between the Russian Academy of Sciences and the SEL (2022) / Scientific electronic library eLIBRARY.RU, 26.05.2022. https://elibrary.ru/projects/rsci/ran_2022.pdf.

Authors

Irina N. Chernova – Senior Researcher, Analytical Research Department, Directorate of State Scientific and Technical Programmes (Russian Federation, 123557, Moscow, Presnensky val Str., 19; e-mail: chernova@fcntp.ru).

Olga V. Cherchenko – Researcher, Analytical Research Department, Directorate of State Scientific and Technical Programmes; Scopus Author ID: 57209975440, ORCID: 0000-0002-2669-0885 (Russian Federation, 123557, Moscow, Presnensky val Str., 19; e-mail: olya.cherchenko@mail.ru).