МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 1. ОРГАНИЗАЦИЯ И МЕТОДИКА ИНФОРМАЦИОННОЙ РАБОТЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 5** **Москва 2022**

ОБЩИЙ РАЗДЕЛ

УДК 004.6–022.59:001:311.311

Е.В. Мельникова

Технология больших данных в наборе методов и средств научного исследования в современной наукометрии*

Рассмотрены проблемы применения технологии больших данных (Big Data) в современном наукометрическом анализе. Важность и актуальность этих проблем обусловлена способностью больших данных существенно повысить эффективность наукометрических исследований путем глубинного анализа мегаобъемов разнородных данных и выявления на этой основе новых смысловых взаимосвязей и закономерностей. Раскрыто основное содержание технологии больших данных; представлен перечень требований к данным, которые позволяют их относить к категории больших данных; отмечено значение этой технологии как передового средства научного исследования в наукометрии. Дана подробная характеристика методов научного исследования. Проанализированы особенности библиометрических, альтметрических, вебометрических и вероятностно-статистических мето-

* Статья подготовлена в рамках исследования по теме FFFU-2021-0002 Государственного задания ВИНИТИ РАН и при поддержке Российского фонда фундаментальных исследований – проект № 20-07-00014.

дов. Подчеркнуто важное место технологии больших данных в современном наборе методов и средств научного исследования в наукометрии.

Ключевые слова: *технология больших данных, Big Data, наукометрия, альтметрия, вебометрия, киберметрия, наукометрические показатели, методы научного исследования, неструктурированные данные*

DOI: 10.36535/0548-0019-2022-05-1

ВВЕДЕНИЕ

Технология больших данных (*Big Data*), разработка которой относится к началу XXI в., применяется для решения конкретных прикладных или теоретических задач в области обработки и анализа огромных объемов информации, накопленной в мире, а также новой информации, формирующейся постоянно, на каждом новом временном отрезке развития человечества и всей нашей планеты в целом. Объемы данных растут экспоненциально. Если в 2020 г. глобальный объем накопленных данных составил порядка 60 зеттабайт[1], то в 2025 г., как ожидается, он уже достигнет уровня в 165-170 зеттабайт [1].

Возможности технологии больших данных позволяют справляться с обработкой растущих объемов информации и оптимизировать различные сферы социальной жизни и отрасли экономики, включая энергетику, логистику, медицину, государственное управление, сферу безопасности, банковскую и биржевую деятельность, область телекоммуникаций, метеорологию и другие сферы. В последние годы эта технология все более широко применяется и в сфере науки, в том числе – в такой научной дисциплине, как наукометрия.

СОВРЕМЕННАЯ НАУКОМЕТРИЯ: МЕТОДЫ И СРЕДСТВА НАУЧНОГО ИССЛЕДОВАНИЯ

Наукометрия – это научная дисциплина, которая на основе метрических исследований изучает «развитие науки как информационного процесса» [2], проявляющего себя через систему научных коммуникаций. Наукометрия использует совокупность методов, основанных на разработке и применении особых числовых показателей (метрик[2]), для решения задач по оценке результативности научной деятельности, исследования процесса развития науки, ее структуры, тенденций и перспектив развития. Важно отметить, что при анализе результативности науки наукометрия может определять как количественные, так и качественные показатели, характеризующие результаты научной деятельности [3]. На международном уровне наработки в области методологии оценки научной деятельности были изложены в Лейденском манифесте для наукометрии [4], принятом в 2014 г.

Рассматривая средства научного исследования в наукометрии, необходимо выделить несколько их категорий: материальные, информационные, математические, логические и языковые. В условиях развития процессов информатизации общества, цифровизации инфо-коммуникационной среды значительно возрастает роль *информационных средств* научного исследования. Вычислительная техника, информационные технологии, системы телекоммуникаций, широко внедряемые в различные сферы общественной жизни, включая науку, коренным образом преобразовывают научную деятельность и становятся для науки средствами познания, осмысления окружающей действительности, т. е. средствами научного исследования.

Значимое место в категории информационных средств научного исследования занимают автоматизированные информационные системы и технологии, среди которых различают традиционные и новые, перспективные технологии. В наукометрии в текущий период к наиболее передовым следует отнести технологию больших данных, которая в составе соответствующих информационных систем будет рассмотрена ниже как важное средство научного исследования.

МЕТОДЫ ИССЛЕДОВАНИЯ

В зависимости от области знания, где методы научного исследования изначально разработаны, в наукометрии они делятся на: библиометрические, альтметрические, вебометрические, вероятностно-статистические и метод профессиональных экспертных оценок. Следует сразу подчеркнуть, что всестороннюю объективную оценку результатов научной деятельности может дать только профессиональная экспертиза; она способна в полной мере учесть содержательные аспекты научной работы. В отличие от метода экспертных оценок, все остальные методы основаны на формальных количественных показателях и служат важными инструментами поддержки принятия решений экспертами [5].

Библиометрические методы. Для наукометрии важно, что в рамках библиометрии разработан набор методов для изучения текстов, документальных потоков и массивов информации, которые наукометрия может использовать для решения своих задач, включая оценку результативности науки, определение «веса» ученого в научном сообществе и др. Среди основных методов библиометрии целесообразно выделить анализ пристатейных ссылок и библиографического сочетания документов [6].

Метод пристатейных ссылок (также называют методом научного цитирования) базируется на идее Ю.Гарфилда об использовании ссылок (количества ссылок) на статьи в научных журналах как средства изучения структуры науки и развития ее направлений, равно как и средства информационного поиска. Метод основан на таком показателе, как индекс

[1] Один зеттабайт (ZB) равен 10^{21} байт.

[2] «Метрика» в переводе с греческого означает «измерение».

научного цитирования, разработанный Ю.Гарфилдом и запущенный в научный оборот в 1964 г. – *Science Citation Index (SCI)* [7]. Сейчас это общепринятый показатель значимости трудов ученого; он представляет собой число ссылок на научные публикации ученого. На базе индекса научного цитирования разработано большое количество производных библиометрических показателей, которые широко используются в современной наукометрии. К ним относятся: индекс Хирша[3] [8], g-индекс (вычисляется как корень из суммарного цитирования работ ученого или сотрудников организации) [9], i-индекс (позволяет выделить ядро наиболее научно активных и востребованных авторов, имеющих наиболее высокий индекс Хирша) [10], импакт-фактор (для научных журналов) и некоторые другие показатели.

Метод библиографического сочетания документов (или библиографической связанности)[4] предусматривает поиск связанных по смыслу документов, авторы которых ссылаются на одни и те же работы [11]. Числом совпадающих ссылок измеряется степень смысловой и тематической связанности документов, что важно для пользователей, осуществляющих поиск публикаций по интересующей их тематике. Библиографическое сочетание образуется между цитируемыми документами. Если два или более документов имеют общие ссылки, то они библиографически связаны. Концепция библиографической связанности помогает исследователям в ретроспективном поиске документов в базах данных.

Альтметрические методы. Альтметрия предлагает метрики, альтернативные традиционным (библиометрическим) метрикам/индексам цитирования. Альтметрики, или альтернативные показатели, формируются на основе обработки информации из социальных сетей, традиционных СМИ, правительственных интернет-порталов, тематических платформ, блогов, профессиональных сетевых сообществ – научной и ненаучной направленности. Альтметрия оценивает результаты научной деятельности не по количеству ссылок на научные статьи, как библиометрия, а по присутствию, упоминанию и использованию статей, имени ученых и их идей в вышеперечисленных информационных ресурсах, по уровню общественного интереса к ним и масштабам реального использования [12]. Альтметрия анализирует их *общественный* вес, предоставляя наукометрии дополнительный материал для оценки результатов деятельности отдельных ученых, реже – коллективов, исследовательских организаций.

Для наукометрии важно, что альтметрики можно использовать как вспомогательный инструмент для оценки уровня внимания участников информационных коммуникаций к ученым, их идеям и публикациям, для определения степени их воздействия на социум, а также для характеристики степени влияния ученых в сообществе. Альтметрики, кроме того, можно применять к научным журналам, книгам, научным конференциям, презентациями т. д. Альтметрики эффективно работают с новыми формами представления научных результатов: теперь это не только статьи в академических журналах, но и видео ролики, записи в блогах, аудиоданные, фотографии и пр. Их обсуждение выходит за границы академического сообщества; к обсуждению подключаются представители вненаучного социума [13]. Традиционные методы оценки не умеют работать с таким многообразием форм и каналов передачи информации, альтметрические – умеют.

Альтметрики формируют достаточно многочисленную группу и подразделяются на несколько видов: просмотры (количество просмотров статей, информации об ученом, научной идее) и скачивания – отражают уровень внимания к результатам научного труда; обсуждения/комментарии (в блогах и на форумах, упоминание в новостях, репосты в социальных сетях) – характеризуют потенциальное влияние ученого/научной идеи в социальной среде; сохранения/закладки – свидетельствуют об общественном интересе и степени воздействия ученого/идеи на пользователей; цитирования (например, ссылка на научную публикацию в экспертных заключениях, в правительственных документах) – отражают воздействие ученого/научной статьи на участников коммуникаций, и некоторые другие.

Следует отметить, что существуют определенные ограничения на использование альтметрических показателей в рамках наукометрии. Так, у части научного сообщества вызывает сомнение достоверность данных альтметрии. Сомнение вызвано тем, что сбор значительного объема информации для дальнейшей обработки альтметрия производит на информационных ресурсах ненаучной направленности. Еще одно ограничение связано с тем, что альтметрики могут отразить степень влияния, например, статьи на участников коммуникаций, но при этом не могут ответить на вопрос, является ли отношение участников к данной статье позитивным или негативным. Эти и некоторые другие особенности альтметрик обусловливают необходимость осторожного отношения к применению альтметрических методов для решения задач наукометрии.

Вебометрические методы. Вебометрия исследует количественные аспекты конструирования и использования информационных ресурсов, структур и технологий в пространстве современного веба – на интернет-сайтах, порталах, интернет-форумах и т.д. Одной из разновидностей веб-пространства является интернет вещей, в котором концентрируются данные датчиков, собирающих и передающих информацию от различных устройств, создавая «вещевые сети». Вебометрия формирует свои методы на базе библиометрических, которые она использует в сетевом режиме. Суть вебометрических методов заключается в

[3] Индекс Хирша ученого равен h, если ученый опубликовал h статей, на каждую из которых сослались как минимум h раз.

[4] Метод библиографической связанности документов (Bibliographic Coupling) был введен американским ученым Кеслером в 1963 г. в работе «Библиографическая связанность между научными документами» (Kessler M.M. Bibliographic coupling between scientific papers // American Documentation. – Wiley-Blackwell. – 1963. – Vol. 14, Issue 1. – P. 10-25). Ю.Гарфилд развил и более глубоко проработал эту концепцию применительно к научным публикациям.

том, что собираемая информация обрабатывается с использованием математических, статистических процедур, включая определение среднеарифметического по каждому вебометрическому индикатору. Получаемые графы, построенные на основе взаимного цитирования, анализируются на базе теории графов и методов анализа социальных сетей *SNA (Social network analysis*) [14], который включает исследование плотности и интенсивности возникающих в ходе социальной коммуникации связей/сетей.

Значимость вебометрических методов для наукометрии заключается в том, что результаты анализа плотности и интенсивности сетевых связей, относящихся к конкретному информационному ресурсу, могут дать характеристику степени воздействия ученого или научной статьи на коммуникационный социум и свидетельствовать об их значимости для участников вэб-коммуникаций. Необходимо также отметить, что особенности анализа данных в сетевом пространстве позволяют решать исследовательские задачи наукометрии с использованием новых, отличных от традиционных средств научного исследования, включая технологию больших данных.

Общее, что объединяет рассмотренные метрические методы научного исследования, состоит в том, что в качестве базы они используют библиометрические показатели и уже на их основе в каждой из сфер разрабатываются свои метрики, которые в наибольшей степени соответствуют потребностям именно этой сферы: в библиометрии – показатели, характеризующие движение потоков документов, в альтметрии – движение преимущественно недокументальных потоков, в вебометрии – смешанных информационных потоков в интернет-пространстве. Кроме того, в киберметрии, например, исследуются информационные потоки в цифровой среде. Помимо этого в научном сообществе в последние годы рассматриваются идеи о выделении в самостоятельную область сетеметрии, занимающейся потоками информации в сетевой среде, медиаметрии [15], работающей с информационными потоками в средствах массовой информации, и некоторых других метрических областей. В каждой из них может осуществляться движение научной информации, которая там обрабатывается существующими в этой среде методами. Поэтому все перечисленные области и их методы исследования представляют практический интерес для наукометрии в решении стоящих перед ней задач.

Вероятностно-статистические методы. Базируются на эмпирических закономерностях, нашедших свое отражение в законах Брэдфорда, Лотки, Ципфа и некоторых других ученых. Законы в равной мере относятся к различным сферам системы научных коммуникаций. Сущность закономерности С. Брэдфорда [16] заключается в следующем. Если журналы расположить в порядке убывания количества помещенных в них статей по определенной теме и полученный список разделить на три зоны с одинаковой численностью статей по этой теме, то количество наименований журналов в зонах растет в геометрической прогрессии (например, 10:100:1000) [13].

Подобная закономерность имеет место и в других сферах системы научных коммуникаций. Так, А. Лотка [17] выявил аналогичный характер распределения ученых по публикационной активности, а Дж. Ципф [18] – распределения слов в тексте по частоте их употребления[5]. В 60-е годы XX в. был установлен присущий этим вероятностно-статистическим закономерностям феномен масштабной инвариантности, т е. свойство сохранять форму уравнений, которые их описывают, при произвольных изменениях объемов информационных массивов и потоков. На основе этого общего свойства – закономерности были объединены в рамках наукометрии в группу вероятностно-статистических методов.

Таким образом, с учетом вышеизложенного можно констатировать, что широкий спектр методов научного исследования в современной наукометрии позволяет проводить мониторинг развития науки, используя все многообразие форм представления информации на ресурсах различных типов, и отражать этот процесс с различных ракурсов, а также разрабатывать и применять оценочные метрические показатели, которые характеризуют результаты научной деятельности и служат инструментом поддержки экспертных групп в оценке результативности науки.

ТЕХНОЛОГИЯ BIG DATA КАК ПЕРЕДОВОЕ СРЕДСТВО НАУЧНОГО ИССЛЕДОВАНИЯ В НАУКОМЕТРИИ

Общая характеристика технологии больших данных

Технология *Big Data* включает особые подходы и методы *глубинного поиска, обработки, хранения и анализа данных.* Глубинный поиск и анализ данных объединяются в понятие *Data Mining*. Функционирование этой технологии [19] осуществляется на основе особо крупных и быстро растущих цифровых баз данных или виртуальных массивов, содержащих структурированные и неструктурированные данные. Традиционные решения перестают работать при высоких значениях объема и скорости поступления данных. Способность того или иного технологического приложения обрабатывать большие массивы данных, поступающих на высоких скоростях, из разнообразных источников и в различных форматах является главным критерием отнесения приложения к технологии больших данных.

Преимущества технологии *Big Data* заключаются в том, что она позволяет раскрывать смысловой потенциал мегамассивов данных за счет поиска ценных закономерностей и фактов путем объединения и глубинного анализа больших объемов данных, которые на первый взгляд не связаны между собой по смыслу. Человеческий мозг не может обнаружить такие закономерности, какие выявляет мощный компьютер в комплексе с технологией больших данных, находя совершенно неожиданные смысловые взаимосвязи.

[5] Если все слова достаточно длинного текста упорядочить по частотности их использования, то частотность *n*-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру *n*.

Следует отметить, что для использования больших данных в наукометрии есть некоторые ограничения. Главные из них – это особенности физического доступа к большим данным, так как такими данными зачастую владеют крупные коммерческие компании, для которых запросы исследователей не являются приоритетными. Более простой путь состоит в использовании данных, доступ к которым имеет меньше ограничений: это большие данные из открытых общественных коммуникаций, включая записи научных дискуссий на социальных порталах, обсуждения в соцсетях отдельных научных идей и ученых, комментарии на интернет-форумах, контент вебстраниц, научные статьи, опубликованные в цифровом формате, оцифрованные архивные документы и т.д.

Формула «Шесть V». Для эффективного применения технологии *Big Data* данные должны обладать рядом качеств в соответствии с формулой «Шесть V». К этим качествам относятся [20]: неоднородность и многообразие данных (*Variety*); общий объём данных, поступающих в обработку (*Volume*); скорость прироста данных (*Velocity*); достоверность данных, обеспечивающая точность результатов их обработки (*Veracity*); изменчивость данных, предполагающая многовариантность их интерпретации в зависимости от контекста (*Variability*); ценность данных (*Value*) – общественная полезность результатов обработки данных, которую технология *Big Data* обеспечивает для пользователей.

Необходимо уточнить, что под неоднородностью и многообразием данных подразумевается, что данные поступают в разных форматах, из различных источников (внутренних и внешних) и разной степени структурированности. Многие наукометрические задачи требуют совместной обработки данных различных форматов и степени структурированности. Это как раз позволяет осуществлять технологии больших данных. По мере роста многообразия данных инструменты *Big Data* становятся все более эффективными.

Структурированные и неструктурированные данные. Это две основные группы данных, которыми может быть в целом представлена вся генерируемая информация. Структурированные данные (СД) составляют 20% от общего объема информации, неструктурированные данные (НД) – 80%. Наукометрия имеет дело с обеими группами данных, у каждой из которых есть свои отличительные черты.

Структурированные данные – это хорошо организованные и точно отформатированные данные, которые в виде букв/текста и чисел хорошо вписываются в связанные строки и столбцы таблиц, подобных, например, файлам *Excel, Google Sheets*. СД часто называют количественными данными; это означает, что их объективный и заранее определенный характер позволяет выражать данные в числах, легко их подсчитывать и измерять. Структурированные данные существуют и хранятся в виде таблиц в формате реляционных баз данных; для их хранения не требуется много места. СД легче поддаются автоматизированной обработке, чем неструктурированные.

Неструктурированные данные – это данные, которые не имеют заранее определенной структуры и хранятся в своих собственных, неструктурированных (исходных) форматах. Существует большое разнообразие форматов неструктурированных данных, что дает наукометрии возможность определять различные характеристики объектов исследования, оценивать их с разных ракурсов. Примерами таких данных являются: текстовые файлы, например, документы в форматах *Word, PDF*; переписка в электронной почте, сообщения в социальных сетях, изображения, видео, аудио файлы, данные датчиков интернета вещей, собирающих и передающих информацию от различных устройств, формируя «вещевые сети», и т.д. Неструктурированные данные также называются качественными данными: это означает, что они имеют субъективный и интерпретирующий характер, их можно разделить на категории в зависимости от характеристик и свойств. В связи с неструктурированностью эти данные не могут быть обработаны и проанализированы с помощью традиционных методов и инструментов. Наиболее распространенный формат существования неструктурированных данных – в рамках нереляционных баз данных. Для хранения таких данных требуется много места; они обычно хранятся в «озерах» данных, репозитариях хранения в необработанных форматах.

Возможности технологии *Big Data* позволяют производить глубинный поиск, обработку и анализ необходимых для наукометрии данных во всем их многообразии – структурированных и неструктурированных данных, данных в различных форматах и из различных источников, перечень которых определяется методологическими потребностями наукометрии.

Технология больших данных в когнитивных и аналитических системах

Технология *Big Data* эффективно применяется в когнитивных[6] и аналитических информационных системах. Обработка структурированной и неструктурированной информации в них и выдача результатов происходит на скоростях, которые намного выше, чем это может делать человек. Объемы данных, с которыми работают системы, также значительно превышают возможности естественного интеллекта. При этом когнитивные системы могут обеспечивать достаточно высокий уровень точности ответов на вопросы, которые пользователи излагают на естественном языке.

Когнитивные системы – это информационные системы, использующие инновационные методы обработки данных, сходные с мыслительными процессами человека. Такие системы обладают способностью автоматизированного самообучения (машинного обучения [21]). Используя технологию *Big Data*, когнитивные системы производят анализ особо крупных и динамически растущих объемов неоднородных данных и выявляют определенные закономерности, которые не мог получить человек/ученый на основе возможностей только естественного интеллекта или на основе применения традиционных программных продуктов об-

[6] «Когнитивный» от лат. Cognitio; означает «обладающий способностью познания, осмысления».

работки данных [22]. Зачастую с помощью технологии больших данных могут выявляться совершенно неожиданные закономерности или характеристики исследуемых объектов, что позволяет находить принципиально новые, более рациональные, менее затратные варианты решения существующих задач.

Аналитические информационные системы – это особый класс компьютерных систем, предназначенных для аналитической обработки и сохранения данных. Системы объединяют информацию, извлекаемую из внутренних баз данных организации и из внешних источников, анализируют ее и хранят как единое целое.

Примерами применения технологии больших данных в аналитических информационных системах являются мировые индексы цитирования, точнее – их аналитические надстройки [13]. В системе индексации и цитирования *Web of Science* функционирует аналитическая надстройка *InCites*, в системе *Scopus* – *SciVal*. В *InCites* сравниваются сводные библиометрические показатели стран и организаций за разные промежутки времени и по разным областям знания. В *SciVal* выполняется кластеризация научных публикаций и графическое представление кластеров в виде «Колеса науки». Надстройки делают аналитические вычисления для поддержки экспертных оценок научных результатов, а также для выявления тенденций и перспектив развития науки.

Аналитическую надстройку, сходную по своим базовым задачам с выше приведенными, имеет российская система индексации и цитирования РИНЦ [23]. Ее аналитическая надстройка *ScienceIndex* позволяет: 1) оценивать результативность научных организаций, отдельных ученых, определять импакт-фактор научных журналов и т.д. и 2) получать общее представление об отраслевом и региональном распределении отечественной науки.

Технология *Big Data* в когнитивных и аналитических системах достаточно эффективно обеспечивает решение задач наукометрического анализа, обрабатывая большие объемы разнородных данных на высоких скоростях и выявляя скрытые смысловые взаимосвязи и закономерности.

Практическое применение технологии больших данных в наукометрическом анализе

Характерной особенностью наукометрических исследований на основе *Big Data* является их междисциплинарность: они выполняются учеными из разных дисциплин, преимущественно – из компьютерных и инженерных наук, медицины, естественных наук (включая физику, химию, биологию и некоторые другие области).

Практическим примером может служить исследование индийских ученых *Keshav Singh* и *Sandeep Kumar*, опубликованное в 2021 г. [24]. В работе представлена методика наукометрического исследования на основе технологии больших данных для систем индексации и цитирования, библиотечных коллекций, научных репозитариев, оцифрованных архивов, баз данных особо крупных размеров. Ученые используют библиометрические данные за последнее десятилетие, полученные из системы *Scopus* в *CSV*-файлах (файлах в формате изображения). Рассматриваются библиометрические особенности документов, проиндексированных системой *Scopus*. Для выявления скрытой информации из загруженного набора данных авторы исследования анализируют плотность библиометрических сетей и интенсивность связей между публикациями, число цитирований и перекрестных цитирований, социтирований и самоцитирований. Ученые делают акцент на сборе и анализе данных о темпах роста числа публикаций, их тематических категориях, географическом распределении, особенностях цитирования. С помощью инструмента *VOSviewer* в исследовании проводится оценка частоты использования ключевых слов. На основе автоматизированного глубинного анализа выявляются высоко цитируемые публикации, наиболее интересные для научного сообщества авторы, авторитетные журналы, влиятельные институты и исследовательские коллаборации. Анализ всего объема данных позволяет определить наиболее «горячие» темы научных исследований в заданный период, выявить новые тенденции в развитии изучаемого исследовательского ландшафта, обозначить наиболее востребованные и перспективные направления будущих исследований и в упреждающем порядке оказать финансовую и организационную поддержку соответствующим областям фундаментальных исследований.

ЗАКЛЮЧЕНИЕ

Таким образом, можно констатировать, что разнообразие методов исследования в современной наукометрии, включая библиометрические, вебометрические и другие, а также неоднородность и многообразие данных, которые обрабатываются в рамках этих методов и обладают перечнем характеристик, соответствующих требованиям технологии больших данных, формируют благоприятную основу для применения этой технологии в наукометрических исследованиях. Такая естественная корреляция между существующими в наукометрии условиями, с одной стороны, и требованиями технологии больших данных – с другой, позволяют сделать позитивный прогноз о возможностях открытия на базе *Big Data* новых значимых наукометрических закономерностей, которые позволят значительно повысить качество исследований в данной научной сфере и увеличить точность наукометрических оценок.

СПИСОК ЛИТЕРАТУРЫ

1. Доклады “DataAge 2020” и “DataAge 2025” аналитической компании International Data Corporation (США) / Корпоративный сайт www.idc.com (дата обращения 23.02.2022).
2. Налимов В.В., Мульченко З.М. Наукометрия: изучение развития науки как информационного процесса. – Москва: Наука. – 1969. – 192 с.
3. Гиляревский Р.С., Мельникова Е.В. Отказ от приоритетности международных индексов научного цитирования при оценке труда ученых в

Китае // Научно-техническая информация. Сер. 1. – 2020. – № 9. – С.19-24; Gilyarevski R.S., Melnikova E.V. Rejection of the Priority of International Science Citation Indeces in the Evaluation of Results of Scientific Activity in China // Scientific and Technical Information Processing. – Springer. – 2020. – Vol. 47, № 3. – P. 194-199.
4. Hicks D., Wouters P., Waltman L., Rijcke S., Rafols I. Bibliometrics: The Leiden Manifesto for research metrics // Nature. – 2015. – Vol. 520. – P. 429-431.
5. Мельникова Е.В.Сравнительный анализ современных подходов России и Китая к оценке результатов научной деятельности // Проблемы национальной стратегии. – 2022. – № 1(70). – С. 153-162.
6. Москалева О.В. Развитие наукометрии: основные вехи // В монографии «Руководство по наукометрии: индикаторы развития науки и технологии» / Под ред. М.А. Акоева. – Екатеринбург: Изд-во Урал. ун-та. – 2021. – 358 с.
7. Гиляревский Р.С., Мельникова Е.В. Институт научной информации США: идеология, преобразования, продукты // Научно-техническая информация. Сер. 1. – 2017. – № 10. – С. 26-31.
8. Hirsch J.E. An index to quantify an individual's scientific research output // Proceedings of the National Academy of Sciences of the United States of America. – 2005. –№ 102(46). – P. 16569-16572.
9. Egghe Leo. Expansion of the field of informetrics: origins and consequences // Information Processing & Management. – 2005. – Vol. 41, № 6. – P. 1311-1316.
10. Prathap G. Hirsch-type indices for ranking institutions' scientific research output // Current Science. – 2006. – Vol. 91, № 11. – P. 1439.
11. Garfild Eugene. "Science citation index" – a new dimension in indexing // Science. – 1964. – Vol. 144. – P. 649-654.
12. Цветкова В.А., Калашникова Г.В. Альтметрические показатели в оценке региональной публикационной активности // Информационные ресурсы России. – 2021. – № 4(182). – С. 20-23.
13. Симоненко Т.В. Наукометрия: объект, предмет, методология // Наукометрия: методология, инструменты, практическое применение: сб. науч. ст. / Ред. А.И. Груша и др. – Минск: Беларуская навука. – 2018. – 343 с.
14. Даденко В.А., Даденко С.В. Метрические исследования как форма анализа научной продуктивности // Аналитика и научное проектирование. – 2019. – № 2. – С. 130-136.
15. Мицкевич А.К. К вопросу о сущности и истоках политической медиаметрии // Философско-гуманитарные науки: сб. науч. ст. / Ред. В.А. Гайсёнок и др. – Минск: Изд-во РИВШ. – 2017. – 420 с.
16. Bredford S.C. Sources of information on specific subjects // Engineering. – 1934. – Vol.137. – P. 85-86.
17. Lotka A.J. The frequency distribution of scientific productivity // Journal of the Washington Academy of Science. – 1926. – № 12. – P. 317-323.
18. Zipf G.K. Selected Studies of the Principle of Relative Frequency in Language. – Cambridge, MA: Harvard University Press. – 1932. – 51 p.
19. Румянцев Д.М. Социальная инженерия и технология Big Data // Шестая межд. науч.-практ. конф. «BIG DATA and Advanced Analytics. BIG DATA и анализ высокого уровня», Минск, Республика Беларусь, 20-21 мая 2020 г. / Сб. материалов. Ч 3. – Минск: Бестпринт. – 2020. – 458 с.
20. Nor Asiakin et al. Exploring big data traits and data quality dimensions for big data analytics application // Springer Science and Business Media / Journal of Big Data. – 2021. – Vol. 8. – P. 1-15.
21. Elshawi R., Sakr S., Talia D. Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service // Big Data Research. – 2018. – Vol. 14. – P. 1-11.
22. Мельникова Е.В. Особенности наполнения научных баз данных для эффективного применения технологии Big Data // Информационные ресурсы России. – 2021. – № 4(182). – С. 6-11.
23. Губа К.С. Большие данные в исследовании науки: новое исследовательское поле // Социологические исследования. – 2021. – № 6. – С. 24-33.
24. Keshav Singh R., Sandeep Kumar S. Emerging trends and global scope of big data analytics: a scientometric analysis // Quality & Quantity. – 2021. – Vol. 55, № 2. – P. 1-26.

Материал поступил в редакцию 09.03.22.

Сведения об авторе

МЕЛЬНИКОВА Елена Владимировна – кандидат технических наук, старший научный сотрудник Отделения теоретических и прикладных проблем информатики ВИНИТИ РАН, Москва
e-mail: verden.mel@yandex.ru

ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ

УДК 004.6:001:002.311.311

Н.А. Мазов, В.Н. Гуреев

Базы данных публикаций научной организации как основа информационных исследований*

Рассматриваются вопросы создания, наполнения и поддержки базы данных, содержащей сведения о научных публикациях сотрудников организации. Подчеркивается необходимость связывания описываемых объектов как внутри базы данных, так и с элементами представления данных во внешних системах, что требуется для функционального поиска и выдачи наиболее полной и точной библиометрической информации. Указывается на необходимость реализации выгрузки данных в заданных форматах для подготовки различных отчетов. Акцентируется внимание на новой функции базы данных публикаций – их популяризации и продвижении в научном информационном пространстве, для чего создаются репозитории и web-реплики. Описан 25-летний опыт ведения базы данных трудов научных сотрудников Института нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН, которая отвечает всем современным требованиям в области наукометрической оценки научной продуктивности и потенциала организации.

Ключевые слова: *база данных, публикации организации, публикационная активность, наукометрия, библиометрия, научная библиотека, информационный поиск*

DOI: 10.36535/0548-0019-2022-05-2

ВВЕДЕНИЕ

Научные публикации и – шире – документы по праву считаются единственным видимым и поэтому измеримым результатом научной деятельности [1, 2]. Это свойство публикаций стало главной причиной их использования как основного объекта различных информационных исследований, а также привело к популярности оценки науки методами библиометрии. Особую значимость количественная оценка приобрела в условиях высокой конкуренции между организациями за финансирование и стремительных темпов увеличения объемов научной информации, когда более затратная экспертная оценка стала менее доступной. Рост объемов научной информации затронул фактически все формы ее функционирования и представления: отмечается ежегодное увеличение количества названий периодических изданий, среднего числа статей в журнале, объема аннотаций и списков литературы, числа узких тематических направлений, стран и коллективов, участвующих в научных исследованиях [3]. Одновременно с этим возросла функциональность и простота использования программных продуктов для проведения библиометрической экспертизы как в виде внешних программных разработок, например, *CiteSpace* и *VOSviewer*, так и в форме встроенного инструментария библиометрических баз данных.

Любые информационные исследования, в том числе наукометрические, требуют наличия фактологической базы, включающей метаописание (реже – полные тексты) публикаций. Научные и образовательные организации при оценке своей научной продукции могут либо полагаться на внешние системы учета библиографической информации, либо вести собственную базу данных публикаций (также называемую коллективной библиографией). Все чаще реализуются наиболее оптимальные подходы к установлению связей и взаимообмену между внутренней и внешними базами данных.

В первой части настоящей статьи представлены преимущества и недостатки в использовании внешних и институциональных баз данных для проведе-

* Исследование выполнено в рамках госзадания ГПНТБ СО РАН (проект № 1021053106841-4-1.2.1;5.8.3).

ния информационных исследований, а во второй – описан 25-летний опыт ведения внутренней базы данных публикаций сотрудников Института нефтегазовой геологии и геофизики им. А.А. Трофимука (ИНГГ) СО РАН, частично отраженный в предыдущих наших работах [4-6].

ПУБЛИКАЦИИ СОТРУДНИКОВ НАУЧНЫХ ОРГАНИЗАЦИЙ ВО ВНЕШНИХ БИБЛИОГРАФИЧЕСКИХ И БИБЛИОМЕТРИЧЕСКИХ СИСТЕМАХ

Внешние системы, такие как *Web of Science*, *Scopus* и ряд тематических баз данных, прежде всего адаптированы к поиску библиографической информации [7] и уже вторично – к анализу больших массивов метаинформации и проведению библиометрических исследований на макроуровнях, к каковым относятся крупные регионы или отдельные дисциплинарные направления. Однако постоянно совершенствующийся инструментарий внешних систем все чаще используется и на средних уровнях – для анализа публикационной активности университетов и научных учреждений.

Преимущества использования библиографической информации о публикациях организации во внешних системах включают следующие возможности:

- проведения информационных, главным образом библиометрических исследований для различных целей, например, сопоставления подразделений организации, выявления перспективных направлений или оптимизации библиотечной подписки. Есть положительные примеры библиометрических исследований с использованием исключительно внешних систем, в частности, на основе совокупного анализа авторских профилей сотрудников университетских факультетов [8];
- использования встроенного инструментария баз данных для отчетных целей, особенно в тех организациях, где слабо организована работа по внутреннему учету публикационной активности. В этом отношении следует отметить систему формирования показателей в Российском индексе научного цитирования (РИНЦ), содержащую готовый (хотя и неисчерпывающий) набор метрик для различных отчетов;
- продвижения результатов научной деятельности в международном научном информационном пространстве при наличии выверенного профиля организации, который информирует заинтересованных лиц о её достижениях, текущих разработках, тематике исследований, кадрах и научном потенциале. Такой информацией могут пользоваться финансирующие организации, надзорные государственные органы, а также научные учреждения соответствующей тематики для создания коллабораций, научные журналисты и обычные граждане.

Перечисленные возможности напрямую зависят от качества представленной во внешних системах информации об учреждении и его сотрудниках. Для российских организаций качество данных в таких системах, как *Web of Science* и *Scopus*, существенно выросло за последнее десятилетие. Совместными усилиями специалистов из научных и образовательных организаций, с одной стороны, и российских представителей служб поддержки этих систем – с другой, были созданы или отредактированы публикационные профили большинства российских учреждений. Несколько раз была инициирована проверка точности сформированных профилей с использованием разработанных в *Clarivate* и *Elsevier* алгоритмов сбора информации о публикациях организаций и методов их упрощенной корректировки. Значимой опцией в условиях постоянного реформирования организаций, их слияния и разделения, стала возможность создания иерархической структуры учреждения в обеих международных системах, которая позволяет проверить как сводные показатели по полному профилю организации, так и отдельные показатели по её филиалам или подразделениям. Особую важность эта функция приобрела в последние годы, отмеченные формированием крупных федеральных исследовательских центров, нередко включающих в свой состав сразу несколько институтов.

В то же время у внешних баз данных как источников учета и анализа информации о публикационной активности организации остаются и существенные недостатки, например:

- проблема однозначной идентификации [7, 9, 10]. И наш собственный опыт [11, 12], и опыт других исследователей [8, 13-15] в редактировании профилей организаций демонстрирует наличие этой проблемы даже при перманентных правках библиографической информации. Неточная атрибуция аффилиаций в случае российских организаций приводит к созданию дублей профилей в *Scopus* в 76 % и к потере 17 % публикаций [13], что намного выше заявленных в *Scopus* значений полноты и точности [16]. Многие публикации научных институтов РАН оказываются отнесены не к ним, а к головной организации «Российская академия наук» [8]. Вклад в проблему идентификации вносят множественные варианты англоязычной передачи названия организации, которое может как переводиться, так и транслитерироваться. Проблемы также возникают при опечатках, сокращении или неполном написании названий организаций. Автоматическое отнесение публикаций к профилю дает сбой в случаях с публикациями филиалов организаций, расположенных в других городах. Решение проблемы могло бы лежать в области создания авторитетных файлов по типу реализуемых ранее ВНИИКИ (ныне СТАНДАРТИНФОРМ), однако подобные инициативы пока не находят широкого применения;
- потеря информации об аффилиации, например, при вхождении публикаций из российских журналов в *Scopus* через базу данных *MedLine*, в которой не ставится задача индексирования аффилиаций, отчего эти сведения теряются и в *Scopus* [17, 18];
- задержка индексирования публикаций внешними системами, что в современных условиях существенно затрудняет подготовку отчетов;
- платный доступ ко многим внешним базам данных [15].

Несмотря на отмеченные подвижки в адаптации внешних систем к библиометрической оценке небольших объектов до сих пор справедливыми остаются утверждения, которые были высказаны еще в

самом начале XXI в. о том, что внутренний учет публикационной активности организаций значительно превосходит по полноте и точности библиографическую информацию во внешних системах [19].

ИНСТИТУЦИОНАЛЬНЫЕ БАЗЫ ДАННЫХ НАУЧНЫХ ПУБЛИКАЦИЙ

Достоверные информационные исследования требуют создания локальных баз данных, поскольку на мезо- и микроуровнях, таких как университетские факультеты, подразделения в научных учреждениях, лаборатории, программы и отдельные исследователи, необходимая для анализа информация будет недоступна во внешних системах в полном объеме [7]. Кроме того, на микроуровнях ошибки во внешних системах становятся наиболее критичными и могут давать значительную погрешность в результатах [7]. Недавний сравнительный анализ представленности публикаций во внутренней и внешних системах показал двукратное превосходство внутренней базы данных по полноте [20].

Проблемы, связанные с институциональными базами данных, характерны для разных стран, поскольку отчетность и библиометрическая оценка распространены по всему миру [21]. Основные требования и структура базы данных для решения информационных задач представлены в зарубежных работах [22, 23], методологические подходы к созданию баз данных публикаций с учетом российской специфики описаны в [24].

Обзор отечественной литературы и собственный опыт общения с коллегами позволяют сделать вывод о недостаточном распространении в российских научных организациях баз данных публикаций, технологически соответствующих современным требованиям, а в ряде учреждений базы данных как таковые отсутствуют [24, 25]. В одних случаях администрации организаций полагаются на внешние системы, недостатки которых были отмечены выше, в других – ведутся обычные ежегодные списки с использованием табличных или даже текстовых процессоров. Такие подходы сопряжены с фактически отсутствующей функциональностью поиска и выдачи информации и большим числом ошибок. Определенной заменой базам данных могут выступать указатели научных трудов сотрудников, которые, тем не менее, имеют самостоятельное значение и часто выпускаются как дополнение к базам данных [26-28].

Наиболее распространенными платформами для создания институциональных баз данных в России стали различные системы автоматизации библиотечной деятельности, в особенности «Ирбис» – разработка Ассоциации ЭБНИТ (http://www.elnit.org/). Поскольку эти системы изначально создавались для перевода в электронный вид библиотечных каталогов, при их адаптации к созданию баз данных публикаций сотрудников научных учреждений сохранилась основная библиотечная концепция реализации полнофункционального поиска. Отметим, что различные CRIS-системы, создаваемые вне библиотек (см. далее), возможно, более функциональны в плане библиометрической оценки и генерации отчетов. Однако заложенные в этих системах иные концептуальные подходы при создании баз данных публикаций не позволяют рассматривать их как самоценные библиографические ресурсы.

Система «Ирбис» представлена в научных и образовательных учреждениях в различных версиях, включая наиболее продвинутые сетевые. Характерная особенность, сделавшая эту систему популярной, – комплексность решаемых ею задач. Так, в Саратовском государственном техническом университете база данных публикаций преподавателей на платформе «Ирбис», во-первых, является основой для прикладных библиометрических задач – рейтинговой оценки сотрудников вуза, составления различных отчетов, статистического анализа публикаций с целью поддержки принимаемых управленческих решений; во-вторых, открытая информация о публикациях выполняет популяризаторскую функцию, повышая видимость и рейтинг образовательной организации [29]. Интересен опыт применения системы «Ирбис» в Красноярском научном центре [30]. Для ускоренного наполнения базы данных трудов сотрудников там используются уже готовые метаописания публикаций, предварительно выгруженные из внешних систем *Web of Science*, *Scopus* и РИНЦ, что дополнительно позволяет отмечать индексацию публикаций в этих системах. Продемонстрированы возможности получения сведений о цитированиях публикаций в международных базах данных посредством программного интерфейса API. Значительные наработки в эксплуатации «Ирбис» демонстрирует ГПНТБ СО РАН, где система используется не только для ведения базы данных трудов сотрудников[1], но и для формирования тематических баз данных, объединяющих публикации сотрудников нескольких организаций определенного региона [31]. Такие базы данных могут использоваться для наукометрического анализа научных направлений в регионе, в том числе для выявления потенциала в проведении определенных исследований, построения сетей сотрудничества и пр.

Кроме систем «Ирбис» в научных центрах используются и другие библиотечные системы, в частности, автоматизированная информационно-библиотечная система МАРК-SQL. На этой платформе, например, ведется база данных публикаций научно-педагогических работников Марийского государственного университета [14]. В Волгоградском государственном техническом университете на примере баз данных трудов сотрудников в формате *MS SQL* продемонстрированы: а) поисковые возможности с применением различных фильтров; б) возможности выгрузки библиографических списков по ГОСТ; в) особенности представления библиографических метаописаний публикаций в сети Интернет, что является актуальной задачей популяризации научных разработок вуза [32].

В ряде организаций имеется опыт разработки уникальных систем, созданных *ad hoc*. Так, собственная программная оболочка в среде *Visual FoxPro* с пред-

[1] http://webirbis.spsl.nsc.ru/irbis64r_01/cgi/cgiirbis_64.exe?C21COM=F&I21DBN=SOTR&P21DBN=SOTR&S21FMT=&S21ALL=&Z21ID=&S21CNR=20

ставлением функционального интерфейса для пользователей реализована сотрудниками кафедры «Информатика» в Московском техническом университете связи и информатики [25]. Особенность формирования базы данных – это участие в ее наполнении самих авторов (аналогичный подход предлагается в [15]). Собственная библиографическая информационно-аналитическая система представлена в Институте проблем информатики РАН. В задачи этой системы входят отслеживание публикационной активности сотрудников, указание на индексацию публикаций во внешних системах и автоматическое формирование библиометрических отчетов по публикациям организации [33]. Особенность базы данных публикаций сотрудников Института проблем передачи информации РАН и ФИЦ «Информатика и управление» РАН заключается в привязке публикаций не к организации, а к конкретному коллективу (в том числе временному), в интересах которого создается эта база данных [34]. В системе реализованы модули поиска по научным проектам и аффилиациям авторов, а также по планированию публикационной активности для реализации проектов. Собственные подходы к формированию баз данных публикаций сотрудников организации представлены библиотекой Института физики твердого тела РАН [35].

В приведенных примерах отмечается интерес создателей баз данных к функции популяризации и продвижения результатов организации во внешнем информационном пространстве. Значимые результаты в этом направлении получены в библиотеке Института вычислительного моделирования СО РАН, где в дополнение к базе данных создан репозиторий публикаций по модели «зеленого» открытого доступа [36, 37]. Как нами уже было отмечено, более широкой видимости достижений организации содействуют и выверенные публикационные профили во внешних системах. Алгоритмы приведения в соответствие внутренней и внешней баз данных на примере РИНЦ представлены сотрудниками Российской государственной библиотеки [28].

Вопросы интеграции гетерогенных данных и управления метаданными для разнородных информационных систем интенсивно изучались сотрудниками Института вычислительных технологий СО РАН [38-40]. Коллективом представлены как ценные теоретические разработки, так и практическая реализация на базе научных организаций Сибирского отделения РАН.

Почти 30-летний опыт создания и поддержки институциональных баз данных накоплен специалистами Библиотеки по естественным наукам (БЕН) РАН (часть коллектива сейчас аффилирована с Межведомственным суперкомпьютерным центром РАН), разработки которых используются во многих научных организациях. Этапы становления баз данных публикаций сотрудников – от ведения картотеки до многофункциональной системы, объединяющей базы данных нескольких организаций, – представлены сотрудниками центральной библиотеки Пущино, подразделения БЕН РАН [27]. Выработаны как общие требования к современной реляционной базе данных и ее архитектуре [24, 41-43], так и подходы к созданию корпоративной системы с объединением баз данных нескольких организаций [44]. Основными требованиями к формированию базы данных заявлены: а) функциональность поиска с использованием логических операторов, которая обеспечивается высокой точностью обработки «эквивалентных» записей; б) формирование связей между объектами описания, что позволяет устанавливать активные ссылки в библиографических описаниях объектов для перехода от автора к его публикациям, организации и источникам, от источника к публикациям и пр.

Несомненный интерес вызывает комплекс технологических решений от сотрудников информационно-аналитического отдела Института катализа СО РАН [15, 20, 45, 46] – разработанная ими CRIS-система SciAct с рядом web-приложений представляет собой многофункциональный ресурс, позволяющий решать широкий спектр задач. Так, система выполняет ставшие уже традиционными функции базы данных – учет публикаций, интеграцию с внешними указателями цитирований, генерацию различных видов отчетов. Примечательна редко реализуемая интеграция в систему профилей авторов, организаций и журналов. Кроме того, система позволяет проводить наукометрические исследования, повышающие качество информационного сопровождения научных исследований. В частности, проведенный на архиве публикаций анализ сроков опубликования статей позволил создать рекомендательный список журналов с указанием медианных сроков опубликования [45]. Кроме учета публикаций следует отметить занесение в систему сведений о преподавательской деятельности сотрудников и их НИОКР.

На основе представленного нами краткого обзора отечественной литературы можно выделить основные характеристики, которым должны удовлетворять институциональные базы данных в текущей системе функционирования научных организаций и их наукометрических оценок:

- наиболее полный учет публикаций сотрудников с использованием широкого набора различных источников – самих авторов, ученого секретаря, внешних библиографических систем, печатных ресурсов по библиотечной подписке [15, 27, 36, 37, 46];
- функциональный поиск научных публикаций организации по разным критериям с использованием различных фильтров [25, 32, 37, 44];
- автоматизированная генерация отчетов / библиографических списков в требуемых форматах [15, 24, 25, 32, 33];
- интеграция с внешними библиографическими системами для минимизации ручного ввода данных, учета цитирований, а также корректировки библиографической информации в этих системах посредством обратной связи или специального инструментария [15, 30, 33, 44, 47];
- оперативное представление данных в удобном пользователям интерфейсе в web-среде [27, 32, 37];
- удаленный доступ к базе данных для комфортной работы пользователей [15, 35].

С использованием институциональных баз данных могут решаться следующие основные задачи:

- проведение информетрических расчетов и исследований на уровнях всей организации (нескольких организаций), ее подразделений и отдельных сотрудников для различных целей: отчетов, оптимизации библиотечной подписки [48], рекомендаций по выбору журналов для опубликования [45], расчета ПРНД [15], отслеживания публикационной активности сотрудников [21, 27], контент-анализа публикаций для выявления научных фронтов и приоритетных направлений [49];
- реализация справочно-информационных функций, предоставляющих пользователю широкий пласт сведений об организации в целом, ее кадровом потенциале, тематическом распределении подразделений, динамике развития тех или иных направлений, научной истории, становлении научных школ и пр. [26];
- реализация популяризаторской функции научных достижений организации, повышающей её рейтинг, в том числе через редактирование на основе внутренней базы данных информации о публикациях организации во внешних системах, а также через создание репозиториев и веб-реплики базы данных [4, 14, 26, 29, 36].

Несмотря на очевидные преимущества институциональных баз данных публикаций перед внешними системами по полноте и точности, в ряде случаев обращение к внешним системам остается неизбежным. Это касается, например, сбора данных о внешних совместителях, членах диссертационных советов из других организаций, участниках междисциплинарных проектов из разных учреждений, информации о которых в организации не будет. Нередки случаи, когда сотрудники указывают свою аффилированность с другой организацией, с которой они также связаны трудовыми отношениями. В ряде учреждений такие публикации не учитываются во внутренней базе данных, хотя для полной оценки продуктивности исследователя эти сведения часто востребованы. Поэтому наиболее сбалансированный подход, который все чаще используется в научных и образовательных организациях, предполагает совместное использование внутренней и внешних баз данных для оптимального и оперативного учета и представления библиографической и библиометрической информации о публикационной активности организации. При этом информация из внешних систем во многом интегрируется в институциональную базу данных и является обязательной для формирования полноценного контента внутренней системы, а на основе внутренней базы данных проводятся корректировки во внешних системах [46].

ПРОГРАММНО-ТЕХНОЛОГИЧЕСКИЙ КОМПЛЕКС МОНИТОРИНГА ПУБЛИКАЦИОННОЙ АКТИВНОСТИ В ИНГГ СО РАН

Общие сведения. База данных трудов сотрудников Института нефтегазовой геологии и геофизики (ИНГГ) СО РАН (http://ibc.ipgg.sbras.ru) функционирует в среде автоматизированной библиотечно-информационной системы CDS-ISIS и содержит полные сведения о публикациях научных сотрудников института со времени его основания – 1957 г. За более чем полувековую историю организация претерпела несколько крупных структурных изменений, включая разделение на две исследовательские организации – Институт нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН и Институт геологии и минералогии им. В.С. Соболева СО РАН. В рутинной работе используется диапазон с 2006 г. – времени последней реорганизации и создания ИНГГ СО РАН в его текущем виде. Кроме публикаций головной организации учитываются публикации трех её иногородних филиалов – Западно-Сибирского, Томского и Ямало-Ненецкого.

Подобно опыту многих других организаций [27, 42] при создании базы данных за основу была взята библиотечная картотека научных публикаций, преобразованная в 1996 г. в машиночитаемый вид. Тогда же данные картотеки были существенно дополнены публикациями из внешних баз данных (РЖ ВИНИТИ – с 1981 г., *Current Contents* – с 1987 г.), а также на основе ежегодных списков публикаций, предоставляемых сотрудниками института в службу ученого секретаря. Для поиска публикаций во внешних системах использовались фамилии сотрудников института на основе данных из отдела кадров, а для поиска в ретроспективу – данные о сотрудниках из библиотечной картотеки. База данных состоит из трех связанных между собой модулей – публикации, авторы и источники публикаций (рис. 1), зарегистрированных в Роспатенте[2].

Модуль описания публикаций содержит более 40 тыс. записей и представляет собой машиночитаемую библиографическую реферативную базу данных. Учитываются следующие типы документов: монографии, диссертации, авторефераты диссертаций, статьи в научных журналах, электронные публикации в Интернете, доклады на конференциях, патентные документы, свидетельства о регистрации программ и баз данных, препринты, а также все публикации сотрудников, индексируемые в *Web of Science, Scopus* и Российском индексе научного цитирования (РИНЦ). С распространением электронных версий публикаций для архивов последних лет почти во всех случаях имеются полные тексты.

Требования к полноте базы данных, содержательности записей и актуальности представленной информации постоянно повышаются, вызывая необходимость внесения все новых атрибутов в описание публикаций, а порой и изменения структуры баз данных [50].

[2] Мазов Н.А., Гуреев В.Н. IPGGTR Труды сотрудников ИНГГ СО РАН (реферативно-полнотекстовая библиография): Свидетельство о государственной регистрации программы для ЭВМ // Свид-во о прогр. 2020621025; RU; № 2020620872, заявл. 10.06.2020, опубл. 19.06.2020, ИНГГ СО РАН; Мазов Н.А., Гуреев В.Н. IPGGAU Авторские идентификационные профили: Свидетельство о государственной регистрации программы для ЭВМ // Свид-во о прогр. 2020621128; RU; № 2020620879, заявл. 10.06.2020, опубл. 02.07.2020, ИНГГ СО РАН.

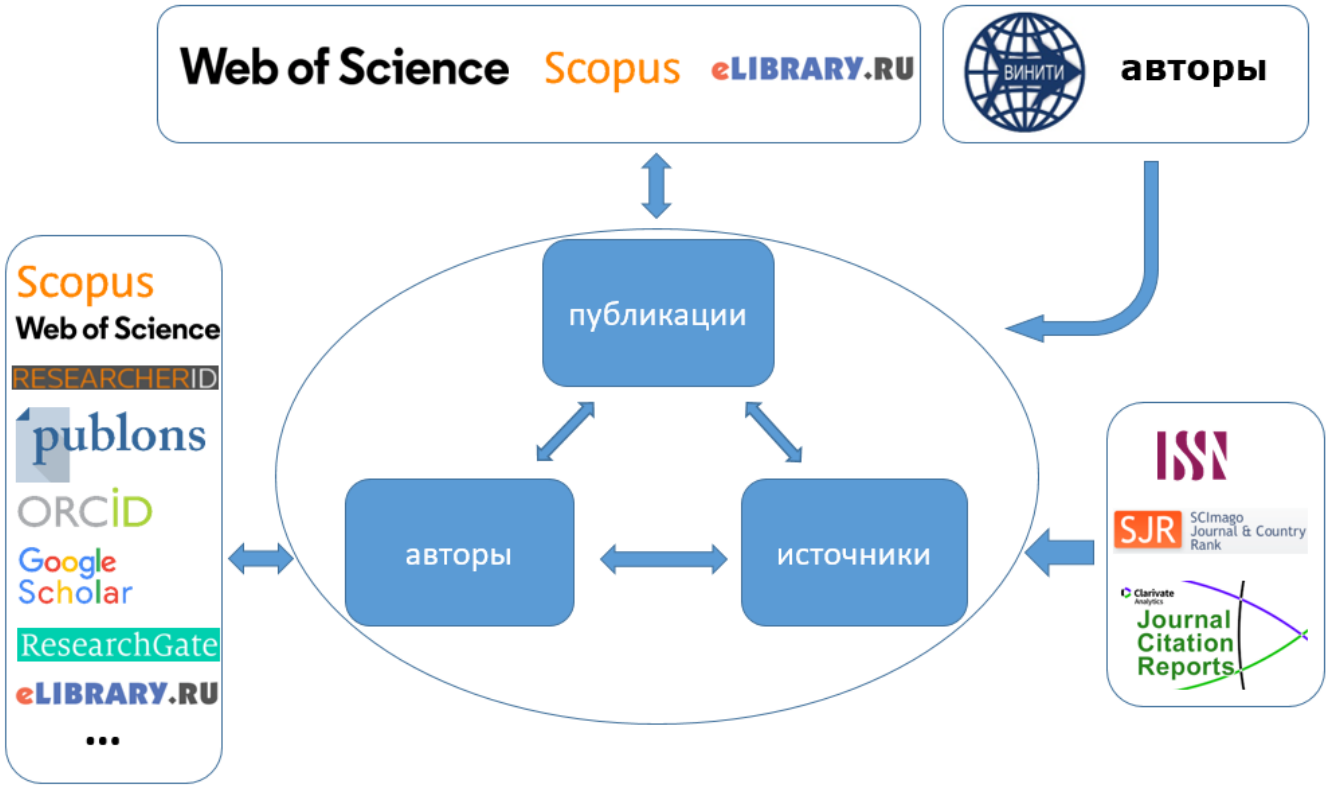

Рис. 1. Комплекс баз данных и взаимосвязи между ними. Источники наполнения

Это связано с возросшей ролью публикационной активности и усложнившейся библиометрической оценкой научных результатов [30, 36, 47], что в свою очередь приводит к росту количества отчетных документов и увеличению их объема. Поэтому к традиционному набору обязательных элементов описания публикаций по ГОСТу в трудах сотрудников ИНГГ СО РАН постепенно добавлялись следующие элементы:

- аннотация на языке оригинала, а для русскоязычных публикаций перевод реферата (при наличии);
- ключевые слова на русском и английском языках;
- сведения об индексировании публикаций в *Web of Science*, *Scopus*, РИНЦ, *Russian Science Citation Index* с указанием идентификаторов, позволяющих перейти на страницу описания публикации во внешних системах;
- идентификатор публикации DOI;
- ссылка URL на страницу с полным текстом публикации (часто пересекается с предыдущим пунктом);
- тираж для монографий;
- номера программ и грантов, по которым финансируются исследования;
- тематические коды ГРНТИ, к которым через таблицы соответствия ВИНИТИ РАН добавлены коды *All Science Journal Classification* из *Scopus*, что позволяет, например, оперативно определять публикационную активность по заданным направлениям работы диссертационных советов;
- число источников в пристатейной библиографии;
- идентификатор внутренней экспертизы по экспортному контролю.

Кроме того, настроено перенаправление между оригинальными и переводными версиями публикаций, что позволяет учитывать их как совместно, так и по отдельности.

Модуль индексации авторов насчитывает более 29 тыс. записей, из них более 1000 уникальных авторских профилей сотрудников института, в каждом из которых учтены все возможные способы передачи имени автора. Модуль связан с общей базой данных сотрудников из кадровой службы организации. Информация о сотрудниках, их принятии на работу, смене лабораторий, увольнении, смене должностных позиций в режиме реального времени поступает в библиотеку, что позволяет с заданной периодичностью обновлять базу данных авторов, а также в случае необходимости отражать изменения во внешних системах.

Характерной особенностью работы с авторскими профилями стал отказ от некоторых правил библиографического описания по ГОСТу, что позволило повысить функциональность поиска. В частности, был реализован ввод всех авторов в порядке следования имени, отчества и фамилии, независимо от их количества и без сокращений. Другим «нарушением» ГОСТа стал ввод всех ответственных лиц (редакторов, переводчиков и пр.) в именительном падеже. Это дало возможность проводить функциональный

поиск по всем персоналиям, внесшим вклад в публикацию, а реализация связки разночтений фамилий авторов (например, при разной транслитерации в зарубежных / переводных публикациях) в один профиль – быстро получать полный перечень публикаций автора. Отметим, что отход от ГОСТа при переходе от картотек на базы данных отмечался и в других библиотеках [27].

Повышенные требования к подготовке библиометрической информации для различных отчетов привели к необходимости отмечать:

- аффилированность с институтом (по умолчанию в базе данных учитываются все публикации сотрудника, независимо от указанной аффилиации);
- общее количество аффилиаций всех соавторов (для упрощения, без их наименований, что достаточно для отчетных целей);
- число аффилиаций, указанное каждым из сотрудников организации;
- сведения о наличии зарубежных соавторов / аффилиаций.

В каждом авторском профиле содержатся данные об идентификаторах сотрудников в различных системах:

- ссылка на страницу сотрудника на сайте организации;
- ссылка на список публикаций автора;
- внутренний идентификатор РИНЦ;
- *SPIN*-код РИНЦ;
- внутренний идентификатор в *Scopus*;
- идентификатор *ORCID*;
- идентификатор *ResearcherID* (перенаправляет на страницу *Publons* с отдельным идентификатором, но может напрямую использоваться в расширенном поиске в *Web of Science*);
- идентификатор *Publons*;
- идентификатор *Google Scholar*;
- идентификатор *ResearchGate*.

Все идентификаторы снабжены гиперссылками, перенаправляющими пользователя на страницу соответствующих систем, а также на страницы сотрудника на сайте института. Эту информацию пользователи могут применять в случае необходимости указания того или иного идентификатора, для прямого доступа в соответствующие системы, а также для составления комплексных запросов во внешние системы, такие как *Scopus* или *Web of Science*. Сводные запросы по авторским идентификаторам, в свою очередь, позволяют настраивать оповещения о новых публикациях. В процессе обработки находятся новые внутренние идентификаторы авторов в *Web of Science*, указанные в адресной строке при поиске по авторским профилям.

Модуль описания источников содержит информацию преимущественно о научных журналах, в которых опубликованы статьи сотрудников института. В настоящее время база данных насчитывает около 500 источников. Для каждого журнала указаны: а) постоянные характеристики, такие как страна издания и коды печатного и электронного ISSN; б) динамические характеристики, включая сведения об индексации в перечне ВАК, списке *RSCI* и квартили журнала по базам данных *Journal Citation Reports* и *SciMago Journal Rank*, которые обновляются в базе данных ежегодно. Соотнесены оригинальные и переводные версии российских журналов.

Процессы формирования базы данных включают поиск информации о публикациях в различных источниках и ее загрузку или ручной ввод во внутреннюю систему. Поиск публикаций основан преимущественно на постоянном мониторинге *Web of Science, Scopus*, РИНЦ и РЖ ВИНИТИ. Информация от сотрудников занимает незначительную долю и затрагивает лишь публикации, отсутствующие в каких-либо индексирующих системах. В *Web of Science* и *Scopus* используются комплексные запросы, включающие возможные вариации в написании названия института, а также запросы по всем сотрудникам организации. В *Scopus* используются авторские идентификаторы *AuthorID*, в *Web of Science* до 2021 г. были запросы по фамилии и инициалам автора. С 2021 г. компанией *Clarivate* в общий доступ выставлены внутренние идентификаторы авторов, по которым стало возможно строить запросы аналогично *Scopus*.

Такой подход реализуется на основе модуля индексации авторов ИНГГ СО РАН и позволяет в любой момент обновить запрос из авторских идентификаторов и настроить новое оповещение. Это позволяет учитывать всю полноту публикаций сотрудников, в том числе не аффилированных с организацией, что может потребоваться для некоторых отчетов. Такой подход реализуется редко: в основном организации учитывают только те публикации, в которых указана организация, и аналогичным образом вносят коррективы во внешние базы данных только по аффилированным с организацией публикациям [28, 51]. Исключениями являются Труды сотрудников ГПНТБ СО РАН, система *SciAct* [20] в Институте катализа СО РАН и некоторые другие.

Следует отметить необходимость периодического ретроспективного поиска публикаций во внешних системах для проставления идентификаторов к публикациям. Это связано с периодической индексацией и в *Web of Science*, и в *Scopus* источников за прежние годы. При этом в *Web of Science* из-за особенностей доступа только к оплаченным временным периодам часть идентификаторов может оказаться недоступной. Так, недоступны в рамках российской национальной подписки публикации до 1975 г., а в 2022 г. из доступа выпали публикации в указателе *ESCI* за период 2015–2016 гг. Ретроспективный поиск требуется и в системе РИНЦ из-за периодической смены идентификаторов публикаций в результате образования дублей, которые, к сожалению, появляются на систематической основе – при вводе публикаций представителями организаций и последующем поступлении метаданных от журналов или из *Web of Science* / *Scopus*.

Процесс ввода метаданных в ИНГГ СО РАН во многом автоматизирован. Аналогично представленному в [30] подходу, в институте используется подготовка данных в *XML*-формате из внешних баз данных для ускоренного их включения во внутреннюю систему. Трудности могут возникать при подготовке данных о русскоязычных публикациях, описанных на

латинице. Проблема становится в настоящее время все более актуальной из-за растущего числа русскоязычных источников и в *Web of Science*, и в *Scopus*. И хотя в *Scopus* в последние два года англоязычное описание статьи дублируется (в части заглавия) оригинальным вариантом, даже в этих случаях необходимо обращение к полному тексту для точной передачи фамилии автора(ов), названия организации и пр.

Опыт использования информации из внутренней базы данных для редактирования профилей организации и авторов во внешних системах мы подробно описывали прежде [5, 11, 12]. Значимость нашей работы заключается в имиджевом продвижении результатов организации, когда сторонние пользователи могут ознакомиться во внешних системах с результативностью организации, ее авторами, тематикой научных разработок и пр., для чего внутренняя база данных может оказаться неподходящей. В общем виде работа в зарубежных базах данных сводится к периодическому поиску оторванных от основного профиля организации публикаций с их последующей привязкой к профилю и устранению дублей авторских профилей. Работа в РИНЦ предусматривает зеркальное отражение структуры организации, поддержку в актуальном состоянии списка авторов, правку их публикационных профилей. Проводимая в последние 10 лет работа позволила в 2–3 раза (в зависимости от библиографической системы) повысить представленность публикаций института во внешних системах.

Функциональность институциональной базы данных. Оригинальное программное обеспечение и реляционный характер трех описанных модулей позволяют решать весь комплекс текущих информационных задач, которые могут стоять перед организацией или ее сотрудниками. Так, реализованы возможности:

- простого, стандартного и расширенного поиска с использованием булевых операторов – по ключевым словам, словам из заглавий, тематическим кодам, видам документов, авторам, источникам, времени выхода публикации;
- получения любого набора библиометрических показателей, которые могут использоваться как в отчетных целях, так и для проведения продвинутых библиометрических исследований по публикациям сотрудников института;
- выгрузки библиографических данных во всевозможных форматах для разных видов отчетов;
- интеграции с внешними базами данных, позволяющими переходить на описания публикаций, источников и авторские профили во внешних системах, а также при необходимости вносить корректировки в эти системы;
- удаленного доступа как к самой базе данных (подробнее о реализации доступа к базам данных *ISIS* в Интернете см. [52, 53]), так и к ее web-реплике с более удобным интерфейсом на сайте института, построенной по типу электронного архива с полными текстами. Отметим, что web-реплика, реализованная отделом информационных технологий ИНГГ СО РАН, повышает видимость записей базы данных в сети Интернет: если сами базы данных, как правило, недоступны для индексирования роботами, то в случае web-реплики все публикации в виде статических html-страниц по стандарту *Dublin core* выставлены для индексирования роботами *Google* и, таким образом, становятся доступными для поиска в *Google Scholar*.

ЗАКЛЮЧЕНИЕ

Создание и ведение институциональной базы данных, как видно из описания процессов работы, требуют постоянного внимания и доработки в ответ на все новые информационные потребности ученых и запросы проверяющих инстанций. Практическое применение подобных систем не исчерпывается утилитарными задачами текущей оценки публикационной активности организации и ее сотрудников. Возможно применение баз данных в качестве полноценных библиографических ресурсов организации и ее подразделений, для оптимизации подписки, разработки рекомендательных систем с автоматизированным подбором нужной пользователю литературы (современный аналог ИРИ) на основе контент-анализа публикаций; исследование приоритетных направлений и научных фронтов, изучение динамики развития тех или иных тематик в организации, ее научной истории, становления научных школ. Все это может быть использовано в корректировке научной политики на уровне отдельных лабораторий, подразделений и организации в целом. Существенную роль внутренние базы данных играют в популяризации научных достижений организации, особенно при создании дублирующих систем в web-среде. Таким образом, современная база данных публикаций сотрудников является значимым, а часто и незаменимым инструментом информационного сопровождения исследований и информированной научной политики.

* * *

Авторы выражают благодарность ведущим библиографам информационно-аналитического центра ИНГГ СО РАН Н.Н. Касаткиной и Е.Н. Томиленко за многолетнее ведение базы данных трудов сотрудников института и техническую реализацию множества необходимых инноваций.

СПИСОК ЛИТЕРАТУРЫ

1. Lazarev V.S. Notion of a document: A center of “gravity attraction” for getting metricians together // Scientometrics. – 1994. – Vol. 30(2–3). – P. 511–516.
2. Лазарев В.С. Библиометрия, наукометрия и информетрия. Часть 2. Объект // Управление наукой: теория и практика. – 2021. – Т. 3(1). – С. 80–105.
3. Thelwall M., Sud P. Scopus 1900–2020: Growth in articles, abstracts, countries, fields, and journals // Quantitative Science Studies. – 2022. – Vol. 3(1). – P. 37-50.
4. Гуреев В.Н., Мазов Н.А. Влияние библиометрических методов на формирование рейтинга научной организации // Труды XV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, элек-

тронные коллекции» «RCDL-2013» (14–17 октября 2013 г., Ярославль). – Ярославль: ЯрГУ, 2013. – С. 118–121.
5. Мазов Н.А., Гуреев В.Н. Библиографическая база данных трудов сотрудников организации: цели, функции, сфера использования в наукометрии // Вестник Дальневосточной государственной научной библиотеки. – 2016. – № 2. – С. 84–87.
6. Мазов Н.А., Гуреев В.Н. Новые методы формирования публикационного профиля научной организации в сети науки // Научные и технические библиотеки. – 2013. – № 12. – С. 42–48.
7. Hood W.W., Wilson C.S. Informetric studies using databases: Opportunities and challenges // Scientometrics. – 2003. – Vol. 58(3). – P. 587–608.
8. Kotsemir M., Shashnov S. Measuring, analysis and visualization of research capacity of university at the level of departments and staff members // Scientometrics. – 2017. – Vol. 112(3). – P. 1659–1689.
9. Jörg B., Höllrigl T., Sicilia M.-A. Entities and identities in research information systems // Proceedings of the 11th International Conference on Current Research Information Systems “E-infrastructures for research and innovation: Linking Information Systems to Improve Scientific Knowledge Production” (6–9 June 2012, Prague, Czech Republic). – Praha: Agentura Action M. – P. 185–194.
10. Мазов Н.А., Гуреев В.Н. Роль единых идентификаторов в информационно-библиографических системах // Научно-техническая информация. Сер. 1. – 2014. – № 9. – С. 32–37; Mazov N.A., Gureev V.N. The role of unique identifiers in bibliographic information systems // Scientific and Technical Information Processing. – 2014. – Vol. 41, № 3. – P. 206–210.
11. Гуреев В.Н., Мазов Н.А. Редактирование профиля организаций в SCOPUS и РИНЦ: сравнение возможностей // Научно-техническая информация. Сер. 1. – 2016. – № 3. – С. 10–22; Gureev V.N., Mazov N.A. Editing organization profiles in Scopus and the RSCI: facilities comparison // Scientific and Technical Information Processing. – 2016. – Vol. 43(1). – P. 66–77.
12. Гуреев В.Н., Мазов Н.А. Идентификация организаций в мультидисциплинарных базах данных: итоги редактирования профилей учреждения в WoS, Scopus и РИНЦ // Труды XVI всероссийской конференции «Распределенные информационные и вычислительные ресурсы. Наука – цифровой экономике» (DICR-2017) (4–7 декабря 2017 г., Новосибирск). – Новосибирск: ИВТ СО РАН, 2017. – С. 450–455.
13. Селиванова И.В., Косяков Д.В., Гуськов А.Е. Влияние ошибок в базе данных Scopus на оценку результативности научных исследований // Научно-техническая информация. Сер. 1. – 2019. – № 9. – С. 25–32; Selivanova I.V., Kosyakov D.V., Guskov A.E. The impact of errors in the scopus database on the research assessment // Scientific and Technical Information Processing. – 2019. – Vol. 46(3). – P. 204–212.
14. Ускова Е.В. Роль библиотеки в работе по повышению публикационной активности и цитируемости трудов научно-педагогических работников МарГУ // Материалы научно-практической конференции «Вузовская библиотека XXI века: перспективы развития» (26 октября 2017 г., Чебоксары). – Чебоксары: Чувашский государственный университет имени И.Н. Ульянова, 2017. – С. 69–76.
15. Альперин Б.Л., Ведягин А.А., Зибарева И.В. SciAct – информационно-аналитическая система Института катализа СО РАН для мониторинга и стимулирования научной деятельности // Труды ГПНТБ СО РАН. – 2015. – № 9. – С. 95–102.
16. Research Metrics Guidebook. – 2018. – URL: https://www.elsevier.com/research-intelligence/resource-library/research-metrics-guidebook (дата обращения: 14.03.2022).
17. Кириллова О.В. Состояние и перспективы представления российских медицинских журналов и публикаций в базе данных Scopus // Вестник экспериментальной и клинической хирургии. – 2014. – Т. 7(1). – С. 10–24.
18. Москалева О.В. Потери публикаций России: почему и как избежать? // 4-я Международная научно-практическая конференция «Научное издание международного – 2015: современные тенденции в мировой практике редактирования, издания и оценки научных публикаций» (26–29 мая 2015 г., Санкт-Петербург). – Санкт-Петербург: Северо-Западный институт управления – филиал РАНХиГС, 2015. – С. 87–91.
19. van Raan A.F.J. The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments // Technikfolgenabschätzung – Theorie und Praxis. – 2003. – Vol. 1(12). – P. 20–29.
20. Альперин Б.Л., Ведягин А.А., Зибарева И.В., Ильина Л.Ю. Система мониторинга и учета научной деятельности SciAct: возможности и перспективы развития // Материалы 21-й Международной конференции и выставки «Информационные технологии, компьютерные системы и издательская продукция для библиотек» (LIBCOM-2017) (20–24 ноября 2017 г., Суздаль). – Москва: ГПНТБ России, 2017. – С. 9.
21. Merceur F., Le Gall M., Salaün A. Bibliometrics: a new feature for institutional repositories // 14th Biennal EURASLIC Meeting “Caught in the “fishing net” of information” (17–20 May 2011, Lyon, France). – Lyon, 2011. – P. 1–21.
22. Mallig N. A relational database for bibliometric analysis // Journal of Informetrics. – 2010. – Vol. 4(4). – P. 564–580.
23. Cobo M.J., López-Herrera A.G., Herrera-Viedma E. A relational database model for science mapping analysis // Acta Polytechnica Hungarica. – 2015. – Vol. 12(6). – P. 43–62.
24. Власова С.А., Каленов Н.Е. Информационная система «Научные труды сотрудников академических учреждений» // Труды XXII Всероссийской научной конференции «Научный сервис в сети Интернет» (21–25 сентября 2020 г., Мос-

ква). – Москва: ИПМ им. М.В. Келдыша, 2020. – С. 152–165.
25. Волков А.И., Воробейчиков Л.А., Сосновиков Г.К. База данных «Публикации преподавателей кафедры» // Методические вопросы преподавания инфокоммуникаций в высшей школе. – 2021. – Т. 10(1). – С. 31–39.
26. Машенцева Л.П. Указатели трудов как индикатор научной активности преподавателей вуза // Материалы IV Международной научно-практической конференции «Модернизация культуры: от культурной политики к власти культуры» (23–24 мая 2016 г., Самара). – Самара: Самарский государственный институт культуры, 2016. – С. 28–33.
27. Захарова С.С., Гуреева Ю.А. Научные публикации: от картотеки трудов до библиографических профилей // Библиосфера. – 2017. – № 2. – С. 85–89.
28. Камышева М.И. Отражение публикаций сотрудников Российской государственной библиотеки в национальной библиографической базе данных научного цитирования (РИНЦ) // Материалы Международной научно-практической конференции «Румянцевские чтения – 2021» (21–23 апреля 2021 г., Москва). – Москва: Пашков дом, 2021. – С. 438–441.
29. Кочетков А.В., Громова Е.В., Ермолаева В.В. Опыт использования библиографической базы данных публикаций сотрудников технического вуза // Научно-техническая информация. Серия 1: Организация и методика информационной работы. – 2015. – № 1. – С. 25–26.
30. Баженов С.Р., Данилин М.В., Рогозникова О.А. Интеграция базы данных публикаций организации с индексами научного цитирования: реализация средствами САБ ИРБИС64 // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: Труды 22-й Международной конференции «Крым-2015» (6–14 июня 2015 г., Судак). – Москва: Изд-во ГПНТБ России, 2015. – С. 1–4.
31. Бусыгина Т.В., Балуткина Н.А., Лаврик О.Л., Мандринина Л.А., Елепов Б.С. Библиографическая база данных трудов сотрудников учреждений СО РАН по нанотехнологиям как инструмент для проведения наукометрических исследований // Информационный Бюллетень РБА. – 2013. – № 66. – С. 192–200.
32. Бахмад Э.А., Курочкина Е.В., Королева И.Ю. Реализация алгоритма обработки данных для предоставления эффективного информационного поиска в базе данных публикаций // Современные проблемы науки и образования. – 2012. – № 2. – С. 1–8.
33. Заикин М.Ю., Обухова О.Л., Соловьев И.В. Библиографическая информационно-аналитическая система ИПИ РАН // Системы и средства информатики. – 2014. – Т. 24(1). – С. 244–259.
34. Иванова А.А., Гладилин С.А., Жуковский А.Е., Плискин Е.Л. База данных для административного учета научных публикаций // Труды ИСА РАН. – 2018. – Спецвыпуск. – С. 83–89.
35. Левченко О.И., Соловьев А.В. Формирование базы данных публикаций сотрудников Института физики твердого тела РАН // Сборник научных трудов «Информационное обеспечение науки: новые технологии» (24–28 августа 2015 г., Таруса). – Москва: БЕН РАН, 2015. – С. 215–221.
36. Ковязина Е.В. Открытый архив и база трудов сотрудников: общность и различие // Материалы международной научно-практической конференции «Корпоративные библиотечные системы: технологии и инновации» (20–27 июня 2016 г., Санкт-Петербург). – Санкт-Петербург: Санкт-Петербургский политехнический университет Петра Великого, 2016. – С. 79–85.
37. Ковязина Е.В. БД трудов сотрудников как средство учета и продвижения научных публикаций // Труды ГПНТБ СО РАН. – 2017. – № 12-2. – С. 336–343.
38. Жижимов О.Л., Федотов А.М., Шокин Ю.И. Технологическая платформа массовой интеграции гетерогенных данных // Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2013. – Т. 11(1). – С. 24–41.
39. Жижимов О.Л., Никульцев В.С., Никульцева Е.В., Федотов А.М., Шокин Ю.И. Технологическая платформа интеграции разнородных распределенных данных ZooSPACE // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: Труды 20-й Юбилейной международной конференции «Крым-2013» (8–16 июня 2013 г., Судак). – Москва: Изд-во ГПНТБ России, 2013. – С. 1–7.
40. Жижимов О.Л., Пестунов И.А., Федотов А.М. Структура сервисов управления метаданными для разнородных информационных систем // Труды XIV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2011) (12–14 октбяря 2011 г., Санкт-Петербург). – Санкт-Петербург, 2011. – С. 1–7.
41. Власова С.А. Автоматизированная система поддержки базы данных научных трудов сотрудников академических учреждений // Информационные ресурсы России. – 2020. – № 5. – С. 29–31.
42. Власова С.А., Каленов Н.Е. Развитие информационной системы регистрации результатов интеллектуальной деятельности сотрудников научного учреждения // Электронные библиотеки. – 2021. – Т. 24(5). – С. 770–793.
43. Vlasova S., Kalenov N. Information system for registering the result of scientific institution employees' intellectual activity // CEUR Workshop Proceedings. – 2020. – Vol. 2784. – P. 283–294.
44. Власова С.А. Автоматизированная система поддержки корпоративной базы данных научных публикаций // Программные продукты, системы и алгоритмы. – 2018. – № 2. – С. 42–46.
45. Альперин Б.Л., Зибарева И.В., Ведягин А.А. Анализ скорости публикации научных статей с использованием CRIS-системы SciAct // Библиосфера. – 2020. – № 1. – С. 83–92.

46. Зибарева И.В., Ведягин А.А., Ильина Л.Ю. Библиометрический учет результативности научной организации в ретро- и перспективе // Труды ГПНТБ СО РАН. – 2017. – № 12-1. – С. 337–346.
47. Баженов С.Р., Рогозникова О.А. Научная публикация как специфический объект описания и требования к базе данных научных публикаций // Материалы Третьего международного профессионального форума «Книга. Культура. Образование. Инновации» – «Крым-2017» (3–11 июня 2017 г., Судак). – Москва: Изд-во ГПНТБ России, 2017. – С. 277–280.
48. Gureyev V.N., Mazov N.A. Detection of information requirements of researchers using bibliometric analyses to identify target journals // Information Technology and Libraries. – 2013. – Vol. 32(4). – P. 66–77.
49. Мазов Н.А., Гуреев В.Н., Глинских В.Н. Библиометрический аспект выявления перспективных направлений в исследовательской организации: на примере наук о Земле // Геофизические технологии. – 2020. – № 3. – С. 4–17.
50. Voronov Y.V., Dmitriev G.I., Zakonnikov E.A. Distinctive features of organizations' scientific potential monitoring databases formation for management decisions support // Proceedings of the 19th International Conference on Soft Computing and Measurements (SCM 2016) (25–27 May 2016, Saint Petersburg, Russia). – Saint Petersburg, 2016. – P. 474–476.
51. Frohlich C., Resler L. Analysis of publications and citations from a geophysics research institute // Journal of the American Society for Information Science and Technology. – 2001. – Vol. 52(9). – P. 701–713.
52. Жижимов О.Л., Мазов Н.А., Фролов А.С. Доступ к базам данных ISIS из Internet и построение распределенной информационной системы // Вычислительные технологии. – 1997. – Т. 2(3). – С. 45–50.
53. Баженов С.Р., Мазов Н.А., Малицкий Н.А., Баженов И.С. Создание программного комплекса доступа из Интернет к базам данных на основе WWW-ISIS // Научные и технические библиотеки. – 1999. – № 2. – С. 47–52.

Материал поступил в редакцию 15.03.22.

Сведения об авторах

МАЗОВ Николай Алексеевич – кандидат технических наук, ведущий научный сотрудник, заведующий информационно-аналитическим центром Института нефтегазовой геологии и геофизики им. академика А.А. Трофимука СО РАН; Государственная публичная научно-техническая библиотека СО РАН, г. Новосибирск
e-mail: MazovNA@ipgg.sbras.ru

ГУРЕЕВ Вадим Николаевич – кандидат педагогических наук, старший научный сотрудник информационно-аналитического центра Института нефтегазовой геологии и геофизики им. академика А.А. Трофимука СО РАН; Государственная публичная научно-техническая библиотека СО РАН, г. Новосибирск
e-mail: GureyevVN@ipgg.sbras.ru

УДК 658:330.356:004.9

П.В. Лимарев

Определение параметров оценки экономической эффективности информационной системы предприятия

Показана необходимость информационной системы на предприятии, обоснованы принципы оценки ее эффективности как инструмента управления. Предложены параметры и методика оценки эффективности информационной системы, которые позволят использовать ее для повышения конкурентоспособности на любом предприятии.

Ключевые слова: *экономическая эффективность, информационные системы, оценка эффективности, бюджет, рентабельность, производственный цикл*

DOI: 10.36535/0548-0019-2022-05-3

Переход к постиндустриальной экономике связан с постепенными, но неизбежными трансформациями во всех сферах жизни общества, в том числе и производственной. Экономика знаний характеризует новый этап развития производства с опорой на ресурсы ее информационного сектора. Использование информационных систем в промышленной политике любого предприятия способствует позитивным производственным изменениям, а результатом становится повышение его конкурентоспособности.

Несмотря на то, что современная экономика характеризуется развитием информационного сектора, значительное число предприятий, в особенности действующих в рамках товарного производства, не придают значения созданию и развитию полноценной информационной системы как эффективного инструмента управления. Во многих случаях предприятие обходится автоматизированной системой бухгалтерского и налогового учёта, а системы планирования ресурсов предприятия ограничиваются автоматизированным учётом товарно-материальных ценностей. Далеко не на каждом предприятии создана и используется система управления взаимоотношениями с клиентами, а система получения и анализа внешних данных и вовсе редкость: чаще всего внешняя информация приобретается бессистемно, некомпетентными лицами и без учёта её качества.

Между тем важнейший фактор успешной деятельности в современных условиях – это информационная система, помогающая предприятию сохранять конкурентоспособность.

Успешность деятельности предприятия подразумевает достижение им эффективности в процессе экономической деятельности, которая складывается из трёх составляющих: производственной (технической) эффективности, аллокационной эффективности (эффективности распределения) и эффективности предпринимательства (рис. 1).

Оценка эффективности каждой из этих составляющих происходит путём сравнения достигнутых результатов с эталонными. Эталонные показатели определяются либо как планируемые (заложенные в бюджет), либо как критические для целей предприятия [1].

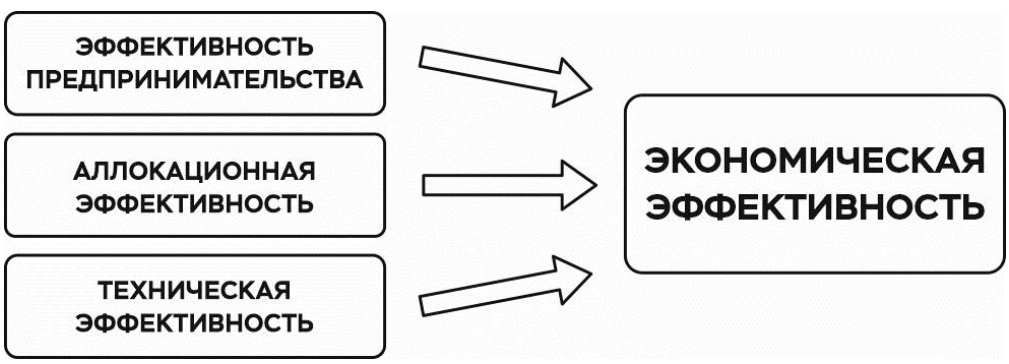

Рис. 1. Составляющие экономической эффективности

Впервые понятие «эффективность» в экономическую теорию ввёл Вильфредо Парето [2]. Модель экономической эффективности, предложенная Парето, подразумевает, что благосостояние общества достигает максимума, а распределение ресурсов становится оптимальным, если любое изменение этого распределения ухудшает благосостояние хотя бы одного субъекта экономической системы. Подробно рассматривали понятие эффективности институционалисты: в 1963 г. Р. Сайерт и Дж. Марч опубликовали труд «Поведенческая теория фирмы» [3], где рассмотрели зависимость эффективности деятельности фирмы от процесса принятия решений; свое видение экономической результативности (на тот момент понятия «эффективность» и «результативность» считались однозначными) предложил Д. Синк в работе «Управление производительностью: планирование, измерение и оценка, контроль и повышение» [4].

Сегодня экономическая наука предлагает несколько устоявшихся определений категории «эффективность». Р.А. Чванов считает, что под эффективностью производства следует понимать эффективность отдачи всех производственных ресурсов в полном объеме, которые использовались или были связаны с получением производственного результата [5]; А.И. Щербаков экономическую эффективность (эффективность производства) определяет как «соотношение полезного результата и затрат факторов производственного процесса» [6].

Методы оценки эффективности информационных систем описывали К.Г. Скрипкин [7], А.Б. Анисифоров и Л.О. Анисифорова [8] и другие исследователи, однако во многих работах информационная система предприятия понимается как набор прикладных программ, кроме того, в ряде случаев эффективность информационной системы не рассматривается с экономической точки зрения; в тех случаях, когда речь идёт об оценке экономической эффективности, параметры оценки в большей степени умозрительные, нежели практические [9, 10].

Информационная система представляет собой инструмент управления предприятием, рассчитанный на обеспечение принятия управленческих решений. Ограничений в формировании информационных систем нет, но традиционно находят применение два вида информационных систем: автоматизированные системы управления предприятием (АСУП), впервые разработанные в СССР в 60-70-х гг. прошлого века (использование АСУП характерно для предприятий, не имеющих проблем со сбытом, чаще всего монопольных и олигопольных, поскольку эта система ориентирована больше на производство, чем на сбыт) и информационные системы по стандартам *Manufacturing Enterprise Solutions Association – MESA* [11], ориентированные на деятельность предприятия в условиях конкурентного рынка. Каждая система имеет свою структуру, свои надстройки и свои преимущества, обсуждение которых в рамках настоящей статьи не планируется.

Наличие информационной системы на предприятии представляет собой потенциальное преимущество. Чтобы это преимущество стало реальным, необходимо использовать информационную систему оптимально, а это зависит от множества факторов: качества информации, оперативности процессов передачи и использования информации, компетентности аналитиков, лиц, принимающих решения и многих других [12].

Деятельность информационной системы предприятия обеспечивает эффективность предпринимательства – компонента экономической эффективности, который позволяет говорить о целесообразности существования коммерческой организации в принципе, и поддерживает два оставшихся компонента – аллокационную и техническую эффективность.

Таким образом для оценки эффективности информационной системы необходимо подобрать параметры, по которым ее использование в рамках деятельности предприятия может быть оценено.

Традиционные показатели эффективности предпринимательства – рост прибыли до налогообложения и рост стоимости бизнеса. Однако оба показателя зависят не только от эффективности использования предприятием информационной системы, но и от множества иных факторов, поэтому для оценки эффективности именно информационной системы следует оценивать параметры, являющиеся элементами общих показателей. Такими параметрами могут служить рентабельность информационной системы (R_{oIS}), исполнение бюджета предприятия и сокращение производственного цикла (рис. 2).

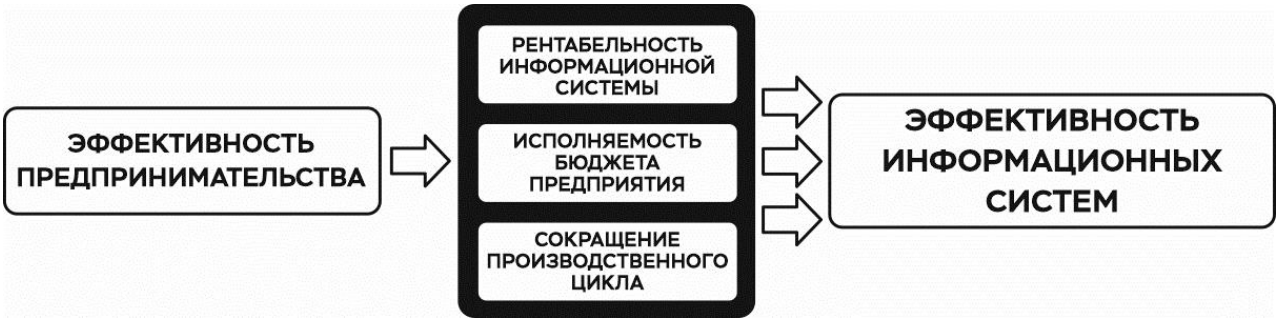

Рис. 2. Параметры оценки эффективности информационной системы

Рентабельность информационной системы определяется как отношение чистой прибыли к расходам на ее функционирование:

$$R_{oIS} = \frac{\Pi}{C_{oIS}}, \quad (1)$$

где: R_{oIS} – рентабельность информационной системы;

П – чистая прибыль;

C_{oIS} – расходы на создание и деятельность информационной системы.

Величина чистой прибыли определяется стандартными методами, а расходы на функционирование информационной системы необходимо определить максимально точно.

Расходы, приходящиеся на создание и деятельность информационной системы, складываются следующим образом:

$$C_{oIS} = EC + PC + EIC, \quad (2)$$

где: C_{oIS} – расходы на создание и деятельность информационной системы;

EC – расходы на оборудование;

PC – расходы на персонал (заработная плата и иные платежи, связанные с заработной платой);

EIC – расходы на приобретение внешней информации.

Кроме того, характерным показателем будет именно изменение рентабельности информационной системы по сравнению с предыдущим отчётным периодом.

Влияние наличия информационной системы на исполнение бюджета предприятия можно оценить по соответствию изменений в исполнении бюджета изменениям в информационной системе при соблюдении следующих условий: прочие параметры, влияющие на исполнение бюджета (трудовая дисциплина, изменение курсов валют, рыночной конъюнктуры и пр.), либо не изменяются, либо их влияние точно определено.

Коэффициент исполняемости бюджета можно выразить отношением плана к фактическому исполнению:

$$R_{BE} = \frac{PB}{AB}, \quad (3)$$

где: R_{BE} – коэффициент исполняемости бюджета;

PB – планируемый бюджет;

AB – фактический бюджет.

Влияние иных факторов на исполняемость бюджета в коэффициенте можно учесть дополнительным множителем:

$$R_{BE}(IS) = R_{BE} \cdot R_{OF}, \quad (4)$$

где: $R_{BE}(IS)$ – коэффициент исполняемости бюджета в зависимости от работы информационной системы;

R_{OF} – коэффициент исполняемости бюджета в зависимости от иных факторов.

Сокращение производственного цикла, так же, как и исполняемость бюджета, может зависеть не только от наличия и результативности деятельности информационной системы, но и от других факторов, которые следует учитывать, и определяется как разница между продолжительностью производственного цикла отчётного периода к продолжительности производственного цикла предыдущего периода:

$$\Delta T(PC) = T(PC)_i - T(PC)_{i-1}, \quad (5)$$

где: $\Delta T(PC)$ – изменение продолжительности производственного цикла;

$T(PC)_i$ – продолжительность производственного цикла в отчётном периоде;

$T(PC)_{i-1}$ – продолжительность производственного цикла в периоде, предшествующем отчётному.

С учётом иных факторов, влияющих на изменение продолжительности производственного цикла, выражение (5) будет выглядеть следующим образом:

$$\Delta T(PC) = T(PC)_i \cdot R(PC)_{OFi} - T(PC) \cdot R(PC)_{OF\,i-1}, \quad (6)$$

где: $R(PC)_{OFi}$ – коэффициент, учитывающий воздействие на изменение продолжительности производственного цикла в отчётном периоде;

$R(PC)_{OFi-1}$ – коэффициент, учитывающий воздействие на изменение продолжительности производственного цикла в периоде, предшествующем отчётному.

Успешная деятельность предприятия в настоящее время базируется на актуальной и достоверной информации. Предприятие, имеющее возможность получения дополнительной информации, связанной как непосредственно со сферой деятельности, так и с любыми аспектами внешней среды, всегда будет иметь конкурентные преимущества перед прочими. Поэтому управление информацией в деятельности промышленного предприятия становится если не главным, то основным направлением в задачах управления [13]. Для получения конкурентных преимуществ необходимо создание полноценной системы информационного менеджмента, которая будет учитывать как внутреннюю, так и внешнюю информацию.

Внедрение и реализация информационной системы на предприятии позволяет приобрести видимые преимущества по сравнению с конкурентами: упрочить положение на рынке, увеличить рыночную долю, добиться увеличения прибыльности.

Предложенные параметры оценки эффективности информационной системы (рентабельность информационной системы, исполнение бюджета предприятия и сокращение производственного цикла) определяются методикой, позволяющей повысить конкурентоспособность предприятия.

СПИСОК ЛИТЕРАТУРЫ

1. Лимарев П.В., Мерзликина Е.М. Инструменты управления экономической эффективно-

стью организации. – Москва: МГУП им. Ивана Фёдорова, 2013. – 112 с.
2. Осипова Е.В. Социологическая система Вильфредо Парето // История буржуазной социологии XIX – начала XX века / под ред. И.С. Кона. – Москва: Наука, 1979. – С. 309-331.
3. Cyert R., March J. A Behavioral Theory of the Firm. – New Jersey: Prentice-Hall Inc., 1963. – 254 p.
4. Синк Д. Управление производительностью: планирование, измерение и оценка, контроль и повышение / пер. с англ. – Москва: Прогресс, 1989. – 528 с.
5. Чванов Р.А. Экономическая эффективность полиграфического производства. Пути повышения. – Москва: Книга, 1981. – 80 с.
6. Щербаков А.И. Совокупная производительность труда и основы её государственного регулирования. Монография. – Москва: Изд-во РАГС, 2004. – 284 с.
7. Скрипкин К. Г. Экономическая эффективность информационных систем в России. – Москва: МАКС Пресс, 2014. – 156 с.
8. Анисифоров А.Б., Анисифорова Л.О. Методики оценки эффективности информационных систем и информационных технологий в бизнесе: учебное пособие. – Санкт-Петербург: СПбГПУ, 2014. – 97 с.
9. Калабина Е.Г., Смирных С.Н. Инструментарий оценки социально-экономической эффективности деятельности организаций государственного сектора экономики. – Екатеринбург: Изд-во УрГЭУ, 2005. – 65 с.
10. Безруков С.Ю., Иващенко Т.И. Методы оценки эффективности ИС предприятия // Научный аспект, 20.02.2020. – URL: https://na-journal.ru/1-2020-informacionnye-tekhnologii/2013-metody-ocenki-effektivnosti-is-predpriyatiya (дата обращения: 11.03.2022).
11. MESA model. – URL: https://mesa.org/topics-resources/mesa-model (дата обращения 11.03.2022).
12. Kotler P. Marketing Management: Analysis, Planning and Control. Englewood Cliffs. – New York: Prentice-Hall, 1967/
13. Лимарев П.В., Лимарева Ю.А. Управление оборотом информации в условиях институциональных ограничений в экономике // Менеджмент в России и за рубежом. – 2015. – № 2. – С. 71-76.

Материал поступил в редакцию 14.03.22.

Сведения об авторе

ЛИМАРЕВ Павел Викторович – кандидат экономических наук, доцент Департамента менеджмента и инноваций Финансового университета при Правительстве Российской Федерации.
e-mail: lavrenty_p@mail.ru

ДОКУМЕНТАЛЬНЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ

УДК 7/9:004.774.25

А.Б. Антопольский

Связанные открытые данные в цифровой гуманитаристике (обзор публикаций)

Предлагается обзор применения технологий связанных открытых данных в зарубежных проектах в сфере цифровой гуманитаристики. Выделяется несколько направлений: преобразование в LOD цифровых коллекций по культуре и искусству; интеграция разнородных данных, связанных с событием; библиографические ресурсы; языковые информационные ресурсы; музыкальные и музыковедческие данные. Особый, наиболее ценный вариант проектов по связанным данным – это технологические разработки и инфраструктурные проекты, создающие основу для национальных систем цифровой гуманитаристики. Подчеркивается роль онтологий в реализации проектов связанных данных.

Ключевые слова: *связанные открытые данные, цифровая гуманитаристика, цифровые коллекции, интеграция данных, библиографические ресурсы, языковые ресурсы, музыкальные данные*

DOI: 10.36535/0548-0019-2022-05-4

ВВЕДЕНИЕ

Технология связанных открытых данных (Linked Open Data – LOD) на платформе Семантической сети, с начала XXI в. стала ведущим направлением для представления научных данных, их интеграции и совместимости, а также коллабораций в этой области.

В области общественных (социальных и гуманитарных) наук технология связанных открытых данных также является наиболее перспективным направлением для интеграции информационных ресурсов. Она рассмотрена автором настоящей статьи в монографии [1].

В настоящей статье содержится обзор и анализ применения этой технологии в зарубежных проектах в сфере цифровой гуманитаристики (DH), подготовленный на основе исследования инфосферы DH, результаты которого опубликованы в работе [2].

Среди проектов в области DH, использующих технологию LOD, можно выделить следующие:

- инфраструктурные и технологические проекты на основе LOD;
- преобразование в LOD цифровых коллекций по культуре и искусству;
- интеграция разнородных данных, связанных с событием;
- библиотечно-библиографические ресурсы в LOD;
- языковые информационные ресурсы в LOD;
- представление в LOD музыкальных и музыковедческих данных.

ИНФРАСТРУКТУРНЫЕ И ТЕХНОЛОГИЧЕСКИЕ ПРОЕКТЫ

В настоящей статье не предполагается дать исчерпывающий анализ инфраструктурных и технологических решений, основанных на связанных открытых данных. Мы приведем лишь некоторые примеры, которые были реально использованы в нескольких проектах DH.

Примером инфраструктурного национального проекта связанных открытых данных для цифровых гуманитарных наук является проект LODI4DH [3], реализуемый в Финляндии.

LODI4DH – это совместная инициатива факультета компьютерных наук Университета Аалто и Центра цифровых гуманитарных наук Хельсинкского университета по созданию централизованных национальных сервисов связанных данных для открытой науки.

LODI4DH основан на большой сети для совместной работы и программном обеспечении, созданном в ходе выполнения длинной череды проектов с 2002 г., в результате которых было создано несколько используемых прототипов инфраструктуры, таких как онтология ONKI служба онтологии Finto в Нацио-

нальной библиотеке Финляндии. Эта служба развернула ONKI и занималась ее дальнейшим развитием на основе SKOS и на платформе Linked Data Finland LDF.fi.

ONKI/Finto и LDF.fi уже имеют широкую пользовательскую базу, что свидетельствует о востребованности инфраструктуры LODI4DH. Приложения на их основе также прошли путь от академических исследований до реального использования. Так, серия смысловых порталов Sampo имеет миллионы пользователей в Интернете. Многие музеи Финляндии, например, Городской музей Эспоо, Консорциум 8 музеев AKSELI и новая национальная система каталогизации KOOKOS, используют онтологии ONKI/Finto. В дополнение к финским проектам существует несколько совместных исследовательских проектов с международными университетами, такими как Оксфорд, Стэнфорд, Колорадо и Пенсильвания, в которых использовались финские службы связанных данных для DH. LODI4DH направлена прежде всего на исследовательскую инфраструктуру DH, но лежащие в ее основе технологии связанных данных и Семантической сети могут использоваться и в других областях исследований, что существенно расширяет потенциал этой инфраструктуры.

Так, данные от сотрудничающих организаций объединяются в общие открытые общедоступные онтологии для: 1) исторических мест и карт, 2) исторических лиц, 3) событий, 4) ключевых понятий и 5) времени. Эти основные онтологии, предоставляемые в виде веб-сервисов, применяются в качестве «семантического клея» при связывании и слиянии данных.

Еще один инфраструктурный проект связанных открытых данных LINCS [4] предназначен для канадских культурных исследований. Этот проект обеспечивает преобразование больших наборов данных во взаимосвязанную машинно-обрабатываемую сеть ресурсов. Исследователям нужна более умная семантическая сеть, которая встраивает смысл в машиночитаемые ссылки, чтобы прояснить различные взаимосвязи понятий и объектов. Технологии связанных открытых данных делают сеть умнее, структурируя и интегрируя данные. LINCS использует эти технологии для увязки канадских исследований и данных о наследии из всего Интернета, преобразуя, соединяя, улучшая и делая доступными ранее гетерогенные и разрозненные наборы данных. Такая увязка обеспечивает пути к новым идеям через сетевое производство знаний как внутри Канады, так и за ее пределами.

Как часть этого проекта осуществляется преобразование метаданных в открытые связанные данные о проектах в сфере культуры [5].

Среди технологических проектов, направленных на развитие LOD, можно указать на метод кодирования связанных данных с помощью JSON (JSON-LD). Соответствующая рекомендация разработана рабочей группой консорциума World Wide Web [6]. Одна из целей JSON-LD заключалась в том, чтобы потребовать от разработчиков как можно меньше усилий для преобразования существующего JSON в JSON-LD. В настоящее время этот стандарт поддерживается Рабочей группой JSON-LD [7].

Проект LodLive [8], разработанный группой итальянских специалистов, демонстрирует использование стандартов связанных данных (RDF, SPARQL) для просмотра ресурсов RDF. Приложение направлено на распространение принципов связанных данных с использованием простого и дружественного интерфейса с многоразовыми методами. Основной принцип, лежащий в основе LodLive, заключается в том, чтобы доказать, что ресурсы, публикуемые в соответствии со стандартом W3C SPARQL, могут быть легко доступны и понятны. Проект предполагает, что подход LodLive может стимулировать государственные администрации и крупных владельцев данных добавлять свои ресурсы в LOD и делиться ими. Можно начинать просмотр, запросив конечную точку для определенного ресурса, или начать с одного из приведенных примеров URI.

Веб-инструмент LodLive разработан для демонстрации стандартов связанных данных, применяемых к просмотру ресурсов RDF с помощью простого интерфейса. LodLive состоит из подключаемого модуля jQuery (lodlive-core.js), карты конфигурации JSON (lodlive-profile.js), HTML-страницы, нескольких изображений (спрайтов) и некоторых других общедоступных плагинов jQuery.

Интересный проект разработан кафедрой цифровых гуманитарных наук Университета Этвёша Лоранда (Венгрия) [9]. Он называется ELTEdata и направлен на организацию просопографических, библиографических и других данных в семантическую сеть и их публикацию. ELTEdata следует структуре Викиданных, связанной с Викиданными семантическими утверждениями и сущностями. Таким образом, ELTEdata может быть интерпретирована как часть Викиданных, хотя и полностью независима от этого ресурса.

Элементы ELTEdata имеют уникальный идентификатор, и каждый семантический оператор может быть описан как пара *свойство–значение*. Язык семантических запросов SPARQL обеспечивает сложный поиск и визуализацию на карте или временной шкале. Эта функция позволяет структурировать большой набор данных и кажется полезной при описании сетей. Слияние с внешними базами данных играет важную роль в формировании семантической библиографии, поэтому актуально сопоставлять некоторые свойства семантических операторов с их вариантами, содержащимися в интерфейсах структурированных метаданных.

К технологическим проектам в области DH можно отнести и проект LIDER *Связанные данные как средство кросс-медиа и многоязычной аналитики контента* [10], определяющий архитектуру, которая создается для анализа многоязычного и мультимедийного контента. Эталонная архитектура LIDER основана на открытых стандартах, существующих и будущих платформах и обеспечивает:

- эталонную модель, определяющую задачи, в которых лингвистические связанные данные могут поддерживать контент-аналитику, и обеспечивающую стандартную декомпозицию этих задач на элементы, которые совместно могли бы решать эти задачи вместе с потоками данных между элементами;

- каталог архитектурных шаблонов, описывающий типы элементов, которые могут быть использованы в вышеупомянутых задачах, типы взаимосвязей между этими элементами, и ограничения на то, как они могут быть использованы.

В проекте описываются различные источники информации для этой эталонной архитектуры, проблемы и препятствия, которые решены лишь частично. Собрана и упорядочена обширная коллекция задач и архитектурных шаблонов NLP в качестве основы для разработки согласованной справочной архитектуры, которая ориентирует первых пользователей данных на основе ссылок из целевых групп лидеров промышленных заинтересованных сторон. Кроме того, эталонная архитектура LIDER включена в контекст деятельности в области связанных данных.

КОЛЛЕКЦИИ КУЛЬТУРНОГО НАСЛЕДИЯ

Центральным направлением цифровой гуманитаристики, вероятно, следует признать создание и обработку разнообразных цифровых коллекций артефактов, относящихся к культурному наследию. Таких проектов в ходе проведенного обследования было выявлено достаточно много. Среди них было несколько, использующих технологии семантической сети и LOD.

Один из наиболее крупных проектов такого рода – это проект ModRef [11] (моделирование, репозиторий, цифровая культура), который объединяет проекты лаборатории Labex PasP [12] из Университета Парижа в Нантере и обеспечивает взаимодействие несколько организаций таких, как:

- MoDyCo (моделирование, динамика, корпус) [13],
- BDIC (Международная библиотека современной документации) [14],
- MAE (Дом археологии и этнологии) [15],
- ArScAn (Археология и античная наука) [16].

Цель ModRef состоит в том, чтобы провести цифровую экспертизу проектов Labex, а также доказать справедливость концепции LOD. В задачу проекта входит разработка моделирования с использованием ссылок.

Проект ModRef должен стимулировать обсуждение вопросов, связанных с миграцией данных в веб-семантику путем создания и использования «хранилищ троек» (коллекций или хранилищ данных RDF-файлов).

Перемещение данных в хранилища троек включает в себя различные этапы:

- подготовка данных (исследование и структурное описание данных),
- семантическое моделирование и сопоставление данных (или сопоставление и выравнивание),
- создание хранилищ троек – перенос данных в хранилища троек,
- публикация и визуализация хранилища троек,
- интерфейс для выполнения sparql-запросов в хранилища троек.

Основные проблемы – это (1) переход от неструктурированных или полуструктурированных данных (блокнот, тексты, html) к структурированным данным (электронные таблицы, реляционные базы данных, XML-файлы), а затем (2) перемещение этих структурированных данных в семантические данные (RDF-файлы) с целью улучшения обмена и открытия новых знаний.

В качестве нормированной онтологии была выбрана CIDOC-CRM [17], поскольку в настоящее время она является справочной для семантического описания данных в музееграфическом или культурном наследии.

CIDOC-CRM доступна в формате OWL, который предоставил Университет Эрлангена-Нюрнберга [18].

Следует отметить, что на сайте ModRef приводится перечень организаций, которые работают над преобразованием метаданных каталогов своих коллекций культурного наследия в связанные открытые данные, часть из них использует в качестве онтологии CIDOC-CRM:

- Британский музей [19],
- Йельский центр британского искусства [20],
- Arches [21] – результат сотрудничества между Институтом сохранения Гетти (GCI) и Всемирным фондом памятников (WMF) по недвижимому культурному наследию,
- Портал Biblissima [22] предоставляет унифицированный доступ к набору цифровых данных о средневековых рукописях, инкунабулах и раннепечатных книгах, выпущенных партнерами консорциума Biblissima. На сайте указано, что предполагается представление всех данных проекта в формате LOD. В настоящее время в этом формате доступны авторитетные файлы системы.

Упомянутые на сайте ModRef известные проекты DH вместо онтологии CIDOC-CRM используют другие системы метаданных, в том числе:

- Онлайн-энциклопедию DBPedia [23] – системы метаданных dbpedia, foaf, umbel, schema.org, dublin core, geo,
- Сервис для хранения, документирования и публикации данных Nakala [24] – системы метаданных foaf, skos, dublin core, vcard,
- Платформа исторической информации Symogih [24] – системы метаданных symogih, example.org, geo.

Потребность в создании связанных открытых данных была рассмотрена в проекте PHAROS Международного консорциума фотоархивов [26]. Это первый шаг на пути к разработке реальной цифровой инфраструктуры фотографических архивов произведений искусства в Европе и Соединенных Штатах Америки. Консорциум организовал сотрудничество между учреждениями, ответственными за четырнадцать фотоархивов с тем, чтобы создать общую платформу для исследований изображений и метаданных.

В качестве примера работы с фотобанками опишем проект преобразования метаданных фотоархива Зери в LOD [27], который включает 290 тыс. фотографий памятников и произведений искусства. Каталогизация хранилища Зери была проведена на основе двух итальянских стандартов метаданных: Scheda F для фотографий и Scheda OA для произведений искусства.

Первый выпуск набора данных Zeri Photo Archive RDF представляет собой значительное подмножество данных, уже доступных на веб-сайте каталога Zeri и для поиска через портал Europeana – это в основном произведения искусства Нового времени (XV–XVI вв.): около 19.000 – сами произведения и более 30 тыс. фотографий, описанных с помощью примерно 11 млн троек RDF.

ИНТЕГРАЦИЯ ДАННЫХ, СВЯЗАННЫХ С СОБЫТИЕМ

Наиболее интересное с точки зрения цифровой гуманитаристики применение технологии LOD заключается в методике интеграции данных и метаданных, посвященных историческим или культурным событиям и объектам культурного наследия, в разных системах хранения, прежде всего в библиотеках, архивах, музеях. Нам уже приходилось делать обзор методов интеграции таких данных [28]. Однако в последние годы в мире было реализовано много новых проектов, интегрирующих разнородные данные на основе технологии LOD. Здесь мы опишем более подробно два таких проекта из различных сфер цифровой гуманитаристики, и кратко – еще несколько.

Первый пример из области социальной психологии. Проект направлен на представление информации по *Стэнфордскому тюремному эксперименту (SPE)* в виде LOD [29] и начинается с отбора по критериям релевантности и неоднородности исходных пунктов (документов), отражающих содержание события (SPE) с различных точек зрения. Для каждого пункта предоставляются основные метаданные, краткое описание и справочная ссылка.

На основе этих пунктов были идентифицированы сущности SPE и отношения между ними. Полученная сложная сеть сущностей затем была очерчена с помощью концептуальной карты (хотя карта является лишь предварительным эскизом, она дает четкий обзор и общее понимание сценария проекта), на основе которой строится модель E/R (сущность/отношение) для SPE. Далее выполняется анализ метаданных и их выравнивание, т.е. приведение к фиксированному списку используемых систем метаданных. Инструментом выравнивания послужил стандарт Dublin Core.

Для каждого объекта разработчики этой модели попытались ответить на вопросы WHO?, WHERE?, WHEN? и WHAT?, определив свойства этих объектов и оформив их в виде стандартных терминов (дескрипторов).

Теоретическая модель подробно предоставляет список свойств, установленных для каждого элемента, и связи между всеми сущностями, включенными в проект, определяя свойства на некоторых авторитетных онлайн-ресурсах таких как, Wikidata, DBpedia, TMDb, Last.FM, Getty Vocabularies, WorldCat и т. д. Эта модель позволила уточнить и расширить модель E/R.

Переходя от теоретической модели к концептуальной модели, разработчики определили наиболее подходящие онтологии, чтобы детально представить элементы проекта. Для подбора онтологий были использованы некоторые инструменты, в частности схема классификации Seeing Standards [30] и портал Linked Open Vocabularies (LOV) [31]. Были выбраны такие наиболее распространенные онтологии:

- *FOAF* (Friend Of A Friend) – онтология, описывающая людей, их деятельность и их отношения с другими людьми и объектами;
- *Schema.org* – стандарт семантической разметки данных в сети, объявленный летом 2011 г. поисковыми системами Google, Bing и Yahoo!;
- *Дублинское ядро.* Термины перечисляют текущий набор словаря Dublin Core, т. е. 15 основных элементов плюс все квалифицированные термины. Он может использоваться для нескольких целей – от простого описания ресурсов до объединения словарей метаданных различных стандартов метаданных и до обеспечения совместимости словарей метаданных в связанном облаке данных и реализациях семантической паутины;
- *CIDOC-CRM*, в настоящее время ISO/CD21127, является основной онтологией, цель которой – обеспечение обмена и интеграции между разнородными источниками информации о культурном наследии, архивами и библиотеками;
- *BIBO* – библиографическая онтология предоставляет основные понятия и правила для описания цитат и библиографических ссылок (т.е. цитат, книг, статей и т.д.);
- *Онтология соглашений* предназначена для моделирования социальных контрактов, которые включают лицензии, законы, контракты, стандарты и метаданные определений;
- *Музыкальная онтология* предоставляет словарь для публикации и связывания широкого спектра данных, связанных с музыкой, в Интернете.

Для всех сущностей в этой модели были построены пространства имен и определены уникальные идентификаторы. Это позволило сформировать RDF-тройки, которые размещены на сайте. Конечная цель проекта – формирование графа знаний события.

Другой пример – интеграция разнородных объектов и сущностей, связанных со сложным историческим событием – «*Фестивалем музыки и искусств Вудсток*» [32]. Главная цель этого проекта – попытка добиться взаимосвязанной системы информации о событии, отталкиваясь от уже существующих в сети ресурсов. Схема выполнения проекта похожа на уже описанную. Сначала были выбраны 10 объектов из разных библиотек, архивов, музеев, посвященных Вудстоку. Для этих объектов на основе естественного языка с помощью интуитивно понятной модели E/R были выделены основные сущности и отношения. Затем разработчики проекта проанализировали стандарты метаданных, которые использовались при описании объектов. Сущности и отношения сравнивались со стандартами метаданных, чтобы выявить наиболее важные аспекты в описании данных.

Когда получили более четкое представление о предметной области, было проведено моделирование данных на абстрактном уровне: построена теоретическая модель, обеспечивающая общее естественно-язычное описание информации о сущностях и отношениях, задействованных в проекте.

Это был промежуточный этап, ведущий к созданию концептуальной модели, способной формализовать описание особенностей данных за счет использования уже существующих онтологических формальных языков. Благодаря различным онтологиям удалось представить и описать данные как формализованные понятия, пригодные для выражения RDF-высказывания в виде триплетов, представленных субъектами, предикатами и объектами. Полученные RDF-высказывания были сериализованы через Turtle, что позволило представить данные в виде триплета URI, выраженного – где это возможно – через онтологические словари.

Были созданы такие взаимосвязи между полученными данными, данными органов власти и других организаций, участвующих в проекте, а также другими ресурсами из онлайн-репозиториев или веб-страниц.

Наконец, все результаты проекта были представлены в виде графа знаний, изображающего контекстуализированную информационную сеть данных, связанных с фестивалем Вудстока, что позволило выявить скрытые отношения, составляющие реальную сущность чего-либо.

Кратко опишем еще несколько проектов такого рода.

WarSamp [33] – это первая крупномасштабная система для обслуживания и публикации LOD о Второй мировой войне: 1) инициирует и способствует крупномасштабной публикации этих данных из распределенных разнородных хранилищ и 2) демонстрирует и предлагает их использование в приложениях и исследованиях DH. Граф знаний содержит более 9 млн высказываний (троек) между элементами данных, включая, например, полный набор из более чем 95 тыс. записей о смертях финских солдат во время Второй мировой войны. В виде RDF-высказываний представлены 160 тыс. сделанных во время войны подлинных фотографий, в том числе 32 тыс. фото исторических мест на исторических картах, 23 тыс. военных дневников воинских частей и 3400 мемуарных статей, написанных ветеранами после войны. Информация в WarSampo поступает из нескольких финских организаций и источников,

WarSampo состоит из двух отдельных компонентов: 1) служба данных WarSampo для машин и 2) семантический портал WarSampo с различными приложениями для пользователей.

Цель проекта *Битва при Ватерлоо в формате LOD* [34] – создание абстрактной модели LOD для описания данных, связанных с битвой при Ватерлоо. Модель должна соотнести событие с личностью «Наполеон Бонапарт», местом «Ватерлоо», датой «1815» и концепцией «Поражение». Последовательность действий: выбор 10 предметов, включающих архивные документы, публикации и артефакты, описывающие идею «Битва при Ватерлоо»; согласование между используемыми учреждениями культурного наследия различными стандартами метаданных, относящимися к людям, месту, дате и концепции. Цель – разработка модели, способной описывать выбранные предметы, и отвечать на вопросы:

- Как описать людей?
- Какая информация о местах?
- В каком формате понятие времени?
- Каково основное содержание объектов?

Модель должна применять существующие формальные системы и онтологии такие, как FOAF, RDFS, SKOS, DC, EAC-CPF.

История экспедиции «Кон-Тики» в формате LOD [35]. При реализации этого проекта была сделана подборка из десяти предметов, связанных с историей Кон-Тики. Для каждого предмета указано название, тип, связанный источник и краткое описание. Сценарий был представлен через модель E/R, содержащую выбранные элементы, сущности и наиболее релевантные отношения между ними. Затем были определены стандарты метаданных, принятые для описания этих предметов.

Чтобы обеспечить функциональную совместимость между стандартами, были определены соответствия между свойствами метаданных. Теоретическая модель – это сценарий, описывающий элементы, связанные метаданные и отношения между ними в абстрактном виде. Модель формально представляет рассматриваемое событие, используя существующие онтологии: графическое изображение было создано, чтобы показать результирующую модель. Выбранные элементы описаны в соответствии с концептуальной моделью с помощью набора таблиц, каждая из которых представляет предметы, предикаты и объекты, подходящие для описания особенностей предметов и основных событий. Сущности и элементы, репрезентативные для данной области, были записаны в виде высказываний RDF, а затем с использованием сериализации Turtle, представлены и объединены. В результате получено графическое изображение знаний о данном событии.

Еще один проект – представление в виде LOD метаданных объектов разных типов (из архивов, музеев, библиотек) по различным аспектам фильма "*Сладкая жизнь*" [36]. По случаю шестидесятилетия выхода этого фильма была реализована модель LOD, ориентированная на интеграцию данных о производстве и распространении этого культового фильма режиссера Федерико Феллини. Данные, когда они изолированы, имеют ограниченный потенциал, их ценность возрастает, когда один или несколько наборов данных, подготовленных и опубликованных независимо и различными субъектами, предлагают возможность интеграции и контакта между ними с целью создания нового общего знания.

Проект История политического диссидента в формате LOD [37] исследует такое событие, как задержание Патрика Заки и его содержание в египетской тюрьме. Цель проекта – моделирование LOD, устанавливающих концепции, предметы, людей, учреждения и места, связанные с задержанием Патрика Заки, чтобы отобразить семантически значимую среду наиболее интегрированным образом.

Было рассмотрено 4 понятийные области:

- эмоции художников, которые создавали визуальные произведения искусства и музыку об этом событии;
- контекстуальная информация об окружающей среде, связанной с Патриком в прошлом и настоящем, об учебном заведении, здании, в котором его задержали;
- условия содержания в тюрьмах Египта на основе докладов Хьюман Райтс Вотч и подкаста итальянского радио RAI;
- юридические аспекты этого события по резолюции, принятой Европейским парламентом, и петиции, подписанной учеными в защиту Заки.

В проекте были изучены возможные технологии, используемые в семантической сети. Упомянем еще два проекта данного класса:

- интеграция ресурсов библиотек, архивов и музеев, связанных с гибелью «Титаника», с использованием LOD [38];

- исследование гендерной проблематики в искусствоведении на основе ARTchives и Викиданных, с использованием технологии LOD [39].

Приведенные примеры убедительно показывают возможность использования технологии связанных открытых данных – LOD для тематически и функционально различных областей цифровой гуманитаристики – DH, когда ставится задача представить по возможности полную информацию о событии с использованием разнородных источников.

БИБЛИОТЕЧНО-БИБЛИОГРАФИЧЕСКИЕ РЕСУРСЫ В ТЕХНОЛОГИИ СВЯЗАННЫХ ОТКРЫТЫХ ДАННЫХ

Специфическим видом ресурсов, которые также можно отнести к области цифровой гуманитаристики, являются библиотечные данные, в том числе библиотечные каталоги, библиографические БД и авторитетные (нормативные) файлы. Библиотечные данные очень удобно представлять в виде связанных открытых данных, и работы в этом направлении в ведущих библиотеках мира ведутся с начала XXI в.

Специально для представления библиотечных данных в LOD разработаны модель Библиотеки Конгресса BIBFRAME [40] и модель Международной федерации библиотечных ассоциаций [41], которые активно применяются различными библиотеками. Общее состояние внедрения LOD в библиотеки можно оценить по опросу, который проводил OCLC [42]. Цель этого опроса – изучение опыта тех библиотек, где реализованы или реализуются проекты/услуги связанных данных. Результаты этого опроса приводятся на сайте OCLC [43], а также в презентации [44]. OCLC и сам ведет исследования технологии LOD [45]. В рамках этого пилотного проекта OCLC и пять партнерских учреждений изучили методы и целесообразность преобразования метаданных в связанные данные для улучшения возможности обнаружения и управления оцифрованными культурными материалами и их описаниям.

В российской информатике известность получил проект О.Л. Лавреновой и ее коллег по представлению в формате LOD классификации знаний, а именно Библиотечно-библиографической классификации (ББК) [46].

Последовательным пропагандистом применения связанных данных в библиотечном деле является О.Н. Жлобинская. Один из последних принадлежащих ей обзоров зарубежного опыта представлен в презентации [47]. Развернутая модель преобразования в LOD библиографических данных из формата RUSMARC, разработанная этим автором, содержится в работе [48].

ЯЗЫКОВЫЕ ИНФОРМАЦИОННЫЕ РЕСУРСЫ

Наиболее очевидным и адекватным объектом для представления в виде связанных данных являются языковые информационные ресурсы, поскольку аналогичные языковые ресурсы в форме тезаурусов и онтологий активно создавались в течение последних десятилетий, а исследование семантических сетей – это традиционная область лингвистической семантики.

Описание технологии LOD применительно к языковым информационным ресурсам и современное состояние языковых связанных открытых данных (LLOD) содержится в монографии Ф.Джимиано и его соавторов [49]. Некоторые действующие проекты в этой области рассмотрены в нашей работе [50]. Поэтому в настоящей статье мы ограничимся перечислением языковых связанных открытых данных и проектов, вошедших в цитированное выше исследование инфосферы цифровой гуманитаристики:

- лингвистические связанные открытые данные [51];
- формат кросс-лингвистических связанных данных [52];
- преобразование лингвистических данных в связанные открытые данные [53];
- кросс-лингвистические связанные данные [54];
- связанные открытые словари [55];
- связанные данные в лингвистике [56];
- связывание корпусных данных способом удобным для NLP [57];
- информация о языках в формате связанных данных [58];
- словари Getty как связанные открытые данные [59].

Кажется, что единственный российский проект этого направления – это проект Д. Усталова по интеграции тезаурусов русского языка в связанные данные [60].

Предложенный список не является исчерпывающим, но дает достаточно полное представление о масштабах применения технологии LOD к компьютерной лингвистике.

МУЗЫКАЛЬНЫЕ И МУЗЫКОВЕДЧЕСКИЕ СВЯЗАННЫЕ ДАННЫЕ

Разработки в области цифровизации музыкальных данных и тем более представление их в виде связанных данных – это относительно новое направление цифровой гуманитаристики и поэтому не получило такого развития, как перечисленные выше. В нашем исследовании было обнаружено два подобных проекта, краткие сведения о которых мы приводим.

Встраивание графов и преобразование музыкальных данных в связанную форму [61]. Цель настоящего проекта состояла в том, чтобы извлечь необработанные данные о музыкальных записях и преобразовать их в связанную форму, из которой можно извлечь знания.

Авторы проекта указывают, что известные ресурсы DBpedia [62] и Викиданные уже предлагают огромное количество сущностей, связанных с музыкой, но для непопулярных музыкальных записей их свойства часто очень ограничены. Использование гораздо более широкого набора данных, специализированных на музыкальных записях, позволяет делать более интересные и точные прогнозы. В области музыкальных записей это может позволить создать систему, которая не только автоматически предсказывает жанры, характеристики и особенности альбомной записи, но и вероятность ее сходства с учетом средних оценок, выставленных пользователем.

Джазовые коллекции в LOD – ресурс для изучения джазовых исполнителей, выступлений и сетей [63].

Проект использует семантические веб-технологии для соединения данных о джазе, включая дискографии, сольные транскрипции и информацию об исполнителях. Проект предполагает интеграцию с другими наборами данных и разработку интерфейса для семантического поиска и визуализации данных.

ЗАКЛЮЧЕНИЕ

Из приведенного обзора связанных открытых данных в цифровой гуманитаристике с очевидностью следует, что в мире сложилась методика создания информационного ресурса для анализа предметной области на основе технологии связанных данных. Чаще всего эта технология применяется при формировании графа знаний применительно к событию, персоне, или тематической коллекции. Для таких направлений цифровой гуманитаристики как библиотечные и языковые данные разработаны специфические модели преобразования данных в форму связанных открытых данных 8 – LOD.

Особый, наиболее ценный вариант работ по связанным данным – это технологические разработки и инфраструктурные проекты, создающие основу для национальных систем цифровой гуманитаристики, как это сделано в Финляндии и в Канаде.

Важная особенность проектов по связанным данным заключается в том, что они опираются на уже созданные онтологии и системы метаданных, которые фиксируют сложившийся в мировой науке консенсус по представлению знаний в определенных областях. Поэтому необходимо создание русскоязычной онтологии гуманитарного знания на основе как российских, так и международных источников, что могло бы стать важным компонентом инфраструктуры российской цифровой гуманитаристики.

СПИСОК ЛИТЕРАТУРЫ

1. Антопольский А.Б., Ефременко Д.В. Инфосфера общественных наук России: монография / под ред. В.А. Цветковой. – Москва; Берлин : Директ-Медиа, 2017. – 676 с. DOI 10.23681/468227.
2. Антопольский А.Б. Инфосфера цифровой гуманитаристики: опыт анализа // Информационные ресурсы России. – 2022. – № 1. – С. 30-38.
3. Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH). – URL: https://seco.cs.aalto.fi/projects/lodi4dh/
4. Linked Infrastructure for Networked Cultural Scholarship. – URL: https://lincsproject.ca/
5. CoDHR. – URL: http://codhr.dh.tamu.edu/2018/04/24/linked-infrastructure-for-networked-cultural-scholarship-lincs
6. JSON-LD 1.1 A JSON-based Serialization for Linked Data. Draft Community Group Report 19 April 2019. – URL: https://json-ld.org/spec/latest/json-ld/
7. JSON-LD Working Group. – URL: https://www.w3.org/2018/json-ld-wg/
8. Live on LodLive. – URL: http://en.lodlive.it/
9. ELTEdata-project. – URL: https://eltedata.elte-dh.hu/wiki/Main_Page
10. LIDER: FP Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe FP7-610782D3. – URL: https://docplayer.net/139432478-Lider-fp-linked-data-as-an-enabler-of-cross-media-and-multilingual-content-analytics-for-enterprises-across-europe.html
11. ModRef Project: Modelling, References, Digital Culture. – URL: http://modref-labexpassespresent.huma-num.fr/ModRef/
12. Labex Past in Present: history, heritage, remembrance – Labex Les Passés dans le Présent: histoire, patrimoine, mémoires. – URL: http://passes-present.eu/
13. MoDyCo. – URL: http://www.modyco.fr/fr/
14. Bibliothèque de la Documentation Internationale Contemporaine. – URL: http://www.bdic.fr/
15. Maison de L'Archéologie et de L'Ethnologie. – URL: http://www.mae.u-paris10.fr/
16. Archéologies et Sciences de l'Antiquité (ArScAn). – URL: http://www.arscan.fr/
17. CIDOC-CRM. – URL: http://www.cidoc-crm.org
18. The Erlangen CRM / OWL. – URL: http://www.erlangen-crm.org/
19. The British Museum. Explore the collection. – URL: https://www.britishmuseum.org/collection
20. Yale Center for British Art. – URL: https://britishart.yale.edu/collections/using-collections/technology/linked-open-data
21. The Getti Conservation Institute Arches Project. – URL: http://www.getty.edu/conservation/our_projects/field_projects/arches
22. Biblissima, the Observatory for Medieval and Renaissance Written Cultural Heritage. – URL: https://biblissima.fr
23. DBPedia. – URL: http://www.dbpedia.org/sparql
24. NAKALA. – URL: https://www.nakala.fr/about
25. SyMoGIH project. – URL: http://www.symogih.org/
26. PHAROS. – URL: http://pharosartresearch.org/
27. The Zeri & LODE project. – URL: http://data.fondazionezeri.unibo.it/
28. Антопольский А.Б. Интеграция библиотечных и архивных информационных систем. – [Б.и.] . – 5 с. – URL: https://rucont.ru/efd/158
29. The Stanford Prison Experiment on LOD. – URL: https://spelod.github.io/#erModel
30. Seeing Standards: A Visualization of the Metadata Universe. – URL: http://jennriley.com/metadatamap/
31. Linked Open Vocabularies (LOV). – URL: https://lov.linkeddata.es/dataset/lov
32. Woodstock Music and Art Festival. – URL: https://woodslod.github.io/woodsLOD/
33. WarSampo Finnish World War II on the Semantic Web. – URL: https://www.sotasampo.fi/en/
34. Battle of the WaterLOD. – URL: https://waterlod.github.io/index.html
35. The Kon-Tiki Expedition. A Linked Open Data project. – URL: https://kontikilod.github.io/KonTiki/
36. La Dolce Vita – Linked Open Data. – URL: https://fellini-lod.github.io/contenuto.html
37. Patrick aLOuD. – URL: https://patrickaloud.github.io/
38. The INKING of RMS Titanic. – URL: https://linkingoftitanic.wixsite.com/linkingtitanic
39. Martrioska. – URL: https://martrioska.github.io/martrioska.html

40. Bibliographic Framework Initiative. – URL: https://www.loc.gov/bibframe/
41. IFLA Library Reference Model. – URL: https://www.librarianshipstudies.com/2020/04/ifla-library-reference-model-lrm.html
42. Online Computer Library Center. – URL: https://www.oclc.org/en/home.html?Redirect =true
43. Linked Data Survey results 6 – Advice from the implementers. – URL: https://hangingtogether.org/linked-data-survey-results-6-advice-from-the-implementers/
44. Smith-Yoshimura K. Linked data implementations – who, what, why? – URL: https://www.oclc.org/content/dam/research/events/2018/smith-yoshimura-lnked-data-implementations-who-whatwhy-SWIB18.pptx
45. Bahnemann G., Carroll M., Clough P., Einaudi M., Ewing Ch., Mixter J., Roy J., Tomren H., Washburn B., Williams E. Transforming Metadata into Linked Data to Improve Digital Collection Discoverability: A CONTENTdm Pilot Project. – Dublin, OH: OCLC Research, 2021. – URL: https://doi.org/10.25333/fzcv-0851.
46. LINKED OPEN DATA. Классификационная система организации знаний. – URL: https://lod.rsl.ru/
47. Жлобинская О.Н. Библиотечные связанные данные: анализ зарубежного опыта . – URL: http://www.nilc.ru/text/NMLBD/NMLBD4.pdf
48. Жлобинская О.Н. Представление библиотечных данных в LOD: возможности и перспективы формата RUSMARC. – URL: http://www.rusmarc.ru/publish/%D0%92%D0%BE%D0%B7%D0%BC%D0%BE%D0%B6%D0%BD%D0%BE%D1%81%D1%82%D0%B8%20%D0%B8%20%D0%BF%D0%B5%D1%80%D1%81%D0%BF%D0%B5%D0%BA%D1%82%D0%B8%D0%B2%D1%8B%20RUSMARC_%D0%A0%D0%BE%D1%81%D1%82%D0%BE%D0%B2.pdf
49. Cimiano Ph., Chiarcos Ch.; McCrae J. P.; Gracia J. Linguistic Linked Data: Representation, Generation and Applications. – Springer International Publishing, 2020.
50. Антопольский А.Б. Лингвистические связанные открытые данные: состояние и перспективы // Научно-техническая информация. Сер. 2. – 2021. – № 8. – С. 28-36. DOI: 10.36535/0548-0027-2021-08-4
51. Linguistic Linked Open Data. – URL: http://linguistic-lod.org/
52. Cross-Linguistic Data Formats. – URL: https://cldf.clld.org/
53. The Prêt-à-LLOD Project . – URL: https://pret-a-llod.github.io/
54. CLLD – Cross-Linguistic Linked Data. – URL: https://clld.org/
55. Linked Open Dictionaries. – URL: http://ionov.me/liodi/
56. Workshop on Linked Data in Linguistics (LDL). – URL: https://www.aclweb.org/anthology/venues/ldl/
57. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. – URL: https://www.researchgate.net/publication/318134320_CoNLL-RDF_Linked_Corpora_Done_ in_an_NLP-Friendly_Way
58. Lexvo.org. – URL: http://www.lexvo.org
59. Getty Vocabularies as Linked Open Data. – URL: http://www.getty.edu/research/tools/vocabularies/lod/index.html#definition
60. Усталов Д.А. Тезаурусы русского языка в виде открытых связанных данных. – URL: https://www.dialog-21.ru/media/1103/ustalovda.pdf .
61. Graph embedding and link prediction starting from a non-linked musical dataset. – URL: https://alerosae.github.io/FromRaw2Linked/
62. DBpedia. Global and Unified Access to Knowledge Graphs. – URL: https://www.dbpedia.org/
63. JazzCats (Jazz Collection of Aggregated Triples). – URL: https://jazzcats.cdhr.anu.edu.au/

Материал поступил в редакцию 17.02.22.

Сведения об авторе

АНТОПОЛЬСКИЙ Александр Борисович – доктор технических наук, профессор, главный научный сотрудник ИНИОН РАН, Москва
e-mail: ale5695@yandex.ru

ВИНИТИ РАН

Центр научно-информационного обслуживания

Информационные услуги, предоставляемые ЦНИО ВИНИТИ РАН:

– проведение тематического поиска и консультации поисковых экспертов;
– подготовка списков научной литературы;
– подбор, копирование полнотекстовых материалов из первоисточников на бумажном носителе и в электронном виде;
– библиометрическая оценка публикационной активности исследователей и научных организаций с использованием российских и зарубежных баз данных;
– информационное обеспечение информационно-аналитической деятельности по подготовке и предоставлению аналитических обзоров и других научных материалов.

ВИНИТИ РАН располагает следующими информационными ресурсами:

– фондом НТЛ, включающим более 2,5 млн. отечественных и иностранных журналов, книг, депонированных рукописей, авторефератов диссертаций и другой научной литературы, ретроспектива – с 1991 года;
– базами данных и Интернет-ресурсами: БД ВИНИТИ (разработка ВИНИТИ), БД SCOPUS, БД Questel (патенты) и другими реферативными ресурсами;
– полнотекстовыми электронными ресурсами (статьи, патенты, материалы конференций).

Ознакомиться с информацией о доступных полнотекстовых и реферативных ресурсах можно на сайте ВИНИТИ РАН www.viniti.ru

К услугам пользователей – **Электронный Каталог ВИНИТИ http://catalog.viniti.ru** и **служба электронной доставки документов.**

Осуществляется платное информационное обслуживание по разовым заказам и на договорной основе с предоставлением всех необходимых финансовых документов.

Проводится индивидуальное обслуживание пользователей в читальном зале ЦНИО ВИНИТИ РАН.

Подробную информацию Вы можете получить:
Адрес: 125190, Россия, г. Москва, ул. Усиевича, 20, ВИНИТИ РАН;
Телефоны: 499-155-42-17, 499-155-42-43;
E-mail: cnio@viniti.ru

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию Реферативный Журнал в электронной форме

РЖ в электронной форме (ЭлРЖ) выпускается по всем разделам естественных, технических и точных наук.

Каждый номер ЭлРЖ является полным аналогом печатного номера РЖ по составу описаний документов, их оформлению и расположению. Он сопровождается оглавлением, указателями.

ЭлРЖ представляет собой информационную систему, снабженную поисковым аппаратом и позволяющую пользователю на персональном компьютере:

- читать номер РЖ, последовательно листая рефераты;
- просматривать рефераты отдельных разделов по оглавлению;
- обращаться к рефератам по указателям авторов, источников, ключевых слов;
- проводить поиск документов по словам и словосочетаниям;
- выводить текст описаний документов во внешний файл.

ЭлРЖ могут быть:

- записаны на DVD-ROM;
- передаваться через FTP-сервер (клиенту предоставляется логин и пароль с доступом к FTP-серверу ВИНИТИ, с которого он скачивает заказанные журналы).

Электронные реферативные журналы можно заказать за текущий год с любого номера, а также за предыдущие годы.

Подробную информацию Вы можете получить:

Адрес: 125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Телефон: 8 499-152-62-11

E-mail: feo@viniti.ru