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. **№ 9** **Москва 2019**

ОБЩИЙ РАЗДЕЛ

УДК 002.63(470)

О.В. Сюнтюренко, Е.Ю. Дмитриева

Государственная система научно-технической информации в структуре задач цифровой экономики*

Показаны задачи Программы «Цифровая экономика Российской Федерации» в развитии информационной и инновационной инфраструктуры. Рассмотрены факторы, определяющие динамику инновационного развития российской экономики. Дан анализ современного состояния ГСНТИ и сформулированы наиболее актуальные и перспективные направления её модернизации. Представлена макроструктура комплекса работ и мероприятий по модернизации системы информационного обеспечения научно-промышленной сферы, факторы, обеспечивающие становление и развитие Национальной информационной системы, и направления разработки и реализации новой научно-информационной политики развития ВИНИТИ РАН. Констатируется, что масштабная задача воссоздания современной отечественной информационной инфраструктуры носит междисциплинарный и надведомственный характер.

Ключевые слова: *информационная инфраструктура, система информационного обеспечения, цифровая среда, направления модернизации, интернет-ресурсы, реферативный журнал, научно-промышленная сфера, аналитическая постобработка, банк данных, информационная безопасность*

* Работа выполнена в рамках проекта РФФИ «Исследование системы классификаторов по науке и технике для разработки смысловой навигации и поиска знаний в информационных сетях», грант № 17-07-00153

ЗАДАЧИ ПРОГРАММЫ «ЦИФРОВАЯ ЭКОНОМИКА РОССИЙСКОЙ ФЕДЕРАЦИИ»

Превращение информации и научного знания в реальную производительную силу изменило характер развития экономики, науки, образования. Традиционный промышленный капитал уступил первенство человеческому капиталу и цифровому капиталу, которые стали превращаться в основные производительные силы современного мира. Инновации становятся важнейшим направлением современного промышленного производства, а интенсификация инновационной деятельности в научно-промышленной сфере – приоритетной задачей экономического развития. Развитие цифровой экономики глобальной информационной инфраструктуры, цифровая трансформация экономического пространства активизируют осознание доминирующей роли информационных ресурсов (ИР) и технологий в процессах мирового экономического и социального развития. По версии международного индекса сетевой готовности, представленной в докладе «Глобальные цифровые технологии» Всемирного экономического форума за 2016 г., Россия значительно отстает от мировых лидеров, занимая «по готовности к цифровой экономике» 41-е место, а по экономическим и цифровым результатам использования цифровых технологий – 38-е место, что объясняется пробелами в нормативной базе для цифровой экономики, недостаточно благоприятной средой для ведения бизнеса и инноваций, низким уровнем применения цифровых технологий, прежде всего в научно-промышленной сфере.

Осознавая важность этой проблематики для развития страны, Правительством РФ была разработана и в июле 2017 г. утверждена Программа «Цифровая экономика Российской Федерации»[1], которая включает пять базовых направлений:

- нормативное регулирование;
- кадры и образование;
- формирование исследовательских компетенций и технических заделов;
- информационная инфраструктура;
- информационная безопасность.

Основная цель направления, касающегося формирования исследовательских компетенций и технологических заделов, – это создание системы поддержки поисковых, прикладных исследований в области цифровой экономики (исследовательской инфраструктуры цифровых платформ), обеспечивающей технологическую независимость каждого из направлений «сквозных» цифровых технологий, конкурентоспособных на глобальном уровне, и национальную безопасность. Здесь следует отметить устойчивую мировую тенденцию сокращения временного лага так называемого инновационного цикла «исследование – разработка – производство». В России создана инфраструктура науки и инноваций, представленная различными научно-исследовательскими институтами, технопарками, бизнес-инкубаторами, которую можно и нужно использовать в целях развития цифровой экономики. В этих условиях модернизация Государственной системы научно-технической информации (ГСНТИ), включающая развитие цифровых ИР, создание качественно новых технологий информационной поддержки наукоемкого производства, как ключевого фактора ускоренного научно-технического и экономического развития, является чрезвычайно важной и актуальной задачей.

Основные цели направления, касающегося информационной инфраструктуры:

- развитие системы российских центров обработки данных, которая должна обеспечивать предоставление государству, бизнесу и гражданам доступных, устойчивых, безопасных и экономически эффективных услуг по хранению и обработке данных и позволит их экспортировать услуги;
- создание эффективной системы сбора, обработки, хранения и предоставления потребителям пространственных данных, которая должна обеспечивать потребности государства, бизнеса и граждан в актуальной и достоверной информации о пространственных объектах;
- организация мониторинга развития цифровой экономики и реализации Программы, разработка и рассмотрение предложений по непрерывному совершенствованию системы управления развитием цифровой экономики, а также установка стандартов и регулирование цифровой экономики.

Основные цели направления, касающегося кадров и образования: создание ключевых условий подготовки кадров для цифровой экономики; совершенствование системы образования, которая должна обеспечивать цифровую экономику компетентными кадрами.

Основная цель направления, касающегося нормативного регулирования: формирование новой регуляторной среды, обеспечивающей благоприятный правовой режим для возникновения и развития современных технологий, а также для экономической деятельности, связанной с их использованием (цифровой экономики).

С учетом задач принятой Программы в рамках настоящей статьи рассматривается актуальная проблематика институционального и технологического развития ГСНТИ и её ведущей организации – Всероссийского института научной и технической информации РАН (ВИНИТИ РАН).

ФАКТОРЫ, ОПРЕДЕЛЯЮЩИЕ ДИНАМИКУ ИННОВАЦИОННОГО РАЗВИТИЯ РОССИЙСКОЙ ЭКОНОМИКИ

Создание инновационной (цифровой) экономики требует разработки эффективной информационной инфраструктуры в сфере научной, научно-технической и инновационной деятельности, адекватной стратегическим установкам построения экономики, основанной на знаниях, развитию сферы науки и инноваций, опережающему росту высокотехнологического сектора, активному продвижению наукоемкой продукции и услуг на мировом рынке. Развитие высокотехнологичного производства и переход к инновационной цифровой экономике сегодня особенно актуальны. Ежегодный спрос на инновационную продукцию РФ

[1] Национальная Программа «Цифровая экономика Российской Федерации»; утверждена распоряжением Правительства Российской Федерации от 28 июля 2017 г. N 1632-р.

составляет лишь $5–7 млрд, а спрос на российское сырье и энергоносители, по оценкам экспертов Торгово-промышленной палаты Российской Федерации, оценивается в размере $500–600 $млрд. На мировом рынке высокотехнологичной продукции удельный вес России составляет около 0,2%, несмотря на инвестиционный рост [1]. Макротенденция развития мирового сообщества – это становление новой мировой «экономики, основанной на знании», в основе которой интеллектуальные и информационные ресурсы, наука и процессы трансфера результатов научных исследований и разработок в продукты, товары и услуги.

Современной российской экономике присущи два существенных, если не сказать важнейших, фактора-детерминанта. Они взаимосвязаны и взаимозависимы.

Во-первых – явная структурно-функциональная недостаточность существующего между фундаментальной наукой и промышленностью «промежуточного слоя», необходимого для создания инновационных продуктов и трансфера технологий. Одна из наиболее актуальных проблем инновационного развития российской экономики связана с существующим разрывом между значительным объемом результатов фундаментальных и прикладных исследований инновационного характера, имеющих потенциал коммерциализации, и фактической способностью и возможностью отечественной промышленности воспринять эти результаты. Такое положение объясняется целым рядом причин финансового, конъюнктурно-экономического, социального и технологического характера. В советский период «промежуточный слой» состоял из отраслевых прикладных НИИ и проектных организаций. В постсоветский период этот «промежуточный слой» практически деградировал, по отдельным направлениям он необратимо деформировался и фактически утратил имевшийся научно-технический потенциал. Сейчас в разных отраслях экономики с различным уровнем эффективности функции «промежуточного слоя» выполняют технопарки, внедренческие центры, венчурные фонды, бизнес-инкубаторы, кластеры, инжиниринговые компании и отдельные сохранившиеся и приспособившиеся к новым условиям НИИ и КБ (в основном в научно-производственных объединениях). В результате новые технологии внедряются крайне слабо. Например, по некоторым оценкам в РФ производится лишь 30% конкурентоспособной продукции, а по оценкам Евросоюза, и того меньше – 5%. По опубликованным данным, в Японии патентуют в 10 раз больше изобретений, чем в РФ. Из этих изобретений в РФ внедряется 0,5%, а 99,5% пылятся на полках.

Во-вторых – несоответствие возможностей существующей национальной информационной инфраструктуры современным требованиям новой российской экономической институциональной среды. Основная системная проблема – темпы развития и потенциал существующей Государственной системы научно-технической информации не позволяют в полной мере удовлетворять растущий спрос и расширяющийся спектр информационных потребностей пользователей из инновационно-промышленной и научно-образовательной сферы народного хозяйства. Системообразующей интегрирующей основой современной информационной инфраструктуры являются цифровые информационные ресурсы и системы. Новые подходы к решению проблем информационного обеспечения и модернизации отечественной информационной инфраструктуры определяются следующими факторами:

- устойчивая тенденция быстрого роста объемов мировых информационных ресурсов;
- доминирующий тренд экспоненциального роста глобальной цифровой среды;
- быстрый рост глобальной сети телекоммуникаций, качественный и количественный рост доступных интернет-ресурсов;
- тенденции сокращения жизненного цикла продукции и сжатия инновационного цикла при одновременном усложнении её разработки и проектирования;
- инновационный вектор развития российской экономики (в условиях санкционного давления).

Следует отметить, что неполное и/или неэффективное использование мировых информационных ресурсов ведет к дублированию исследований и разработок, а в промышленной сфере – к перерасходу энергии, материальных ресурсов, овеществленного живого труда. Удельный вес повторно предлагаемых решений в различных областях научно-технической деятельности весьма значителен – ежегодные прямые потери измеряются в промышленно развитых странах многими миллиардами долларов.

Очевидно, что проблема преодоления инерционного развития национальной информационной системы требует ее модернизации на основе новых концептуальных подходов и технологий.

КОНЦЕПТУАЛЬНЫЕ ЗАДАЧИ МОДЕРНИЗАЦИИ ГОСУДАРСТВЕННОЙ СИСТЕМЫ НАУЧНО-ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ

Научно-техническая информация – это один из главных и дорогостоящих элементов государственных ресурсов. Поэтому упорядочение усилий многих организаций, создающих этот специфический вид ресурсов за счет средств государственного бюджета, является чрезвычайно важной, приоритетной задачей, решение которой должно быть найдено в рамках воссоздания Государственной системы научно-технической информации. В постсоветский период государственным регулятором ГСНТИ, после Госкомитета СМ СССР по науке и технике (ГКНТ), последовательно были: Министерство науки, высшей школы и технической политики; Министерство образования и науки; Министерство науки и высшего образования (в настоящее время). Существенное снижение эффективности и частичная деградация ГСНТИ после 1991 г. в значительной степени объясняются двумя основными факторами. ***Во-первых*** – это процессы общей стагнации отечественного научно-промышленного сегмента в последние двадцать пять лет. Прекратили свое существование центральные отраслевые органы НТИ (ЦООНТИ), включая такие значимые, как «Информэлектро», ЦНИИ «Электроника», НИИ ЭКОС и др. Фактически деградировала сеть региональных

центров НТИ «Росинформресурс». Прекратила свою деятельность система передачи научно-технических достижений оборонных отраслей промышленности в гражданские отрасли народного хозяйства. Резко сократились заинтересованность и платежеспособный спрос на научно-техническую информацию большинства отраслей промышленности. ***Во-вторых*** – это фактическая утрата контроля регулятора (со стороны государства – это федеральные министерства) за функционированием ГСНТИ, прежде всего в сфере материально-финансового обеспечения и целеполагания. С некоторыми оговорками к регуляторам можно отнести и Российскую академию наук, в состав которой входит ВИНИТИ РАН – ведущая организация ГСНТИ.

Базовой правовой основой современной ГСНТИ является «Положение о государственной системе научно-технической информации», утвержденное постановлением Правительства Российской Федерации от 24 июля 1997 г. № 950.

В целом жизнестойкость ГСНТИ в значительной степени объясняется устойчивостью основных информационных центров и научно-технических библиотек страны (ВИНИТИ, ИНИОН, БАН, ГПНТБ, БЕН РАН, ГПНТБ СО РАН, Всероссийская патентно-техническая библиотека и др.), которые, несмотря на все существующие детерминанты, реализуют информационное обеспечение научно-промышленной и образовательной сферы [2].

Несмотря на значительные проблемы социально-экономического развития последнего десятилетия, мы считаем необходимым, хотя бы кратко, отметить появление новых элементов отечественной информационной инфраструктуры:

- Федеральный портал по научной и инновационной деятельности (www.sci-innov.ru), функционирует с 2005 г. Информационным наполнением занимается разработчик Портала – НИИ информационных технологий и телекоммуникаций «Информика»;
- Информационная система Российского фонда фундаментальных исследований (РФФИ) (http://www.rfbr.ru/rffi/ru) содержит заявки и научные отчеты по выполненным фундаментальным и прикладным исследованиям. По экспертным оценкам около 10% завершенных исследований имеют перспективу дальнейшей коммерциализации;
- Единая государственная информационная система учета результатов научно-исследовательских, опытно-конструкторских и технологических работ гражданского назначения, выполненных за счет средств федерального бюджета (http://www.rosrid.ru) функционирует с 2014 г.;
- Информационная система ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технического комплекса России на 2014-2020 годы» (www.fcntp.ru). Программа обеспечивает поддержку перспективных исследований и разработок на всех стадиях инновационного цикла: от генерации знаний – через разработку технологий – к коммерциализации;
- Научная электронная библиотека (eLibrary), интегрированная с информационно-аналитической системой РИНЦ (Российский индекс научного цитирования), – это негосударственная система, но информационное наполнение БнД осуществлялось до настоящего времени на средства федерального бюджета (гранты РФФИ и ФЦП);
- Государственная информационная система промышленности (http://www.gisp.gov.ru); функционирует с 2017 г. Содержит информацию: о прогнозах выпуска основных видов промышленной продукции, её характеристику, а также об объеме импорта промышленной продукции; об использовании ресурсосберегающих технологий и возобновляемых источников энергии в процессе промышленной деятельности; о государственных и региональных программах, разрабатываемых с целью формирования и реализации промышленной политики;
- Национальная электронная библиотека (НЭБ) призвана предоставлять доступ к оцифрованным документам в российских библиотеках, музеях, архивах (www.нэб.ru);
- Научные социальные сети (например, «Ученые России») (http://www.scipeople.ru);
- Пилотный проект КиберЛенинка – научная электронная библиотека, построенная на парадигме открытой науки (*Open Science*), где по лицензионному договору размещено около 500 научных журналов открытого доступа.

Мировые тенденции информатизации и новые задачи, поставленные государством по повышению эффективности информационного обеспечения отечественной науки и промышленности, позволяют сформулировать наиболее актуальные и перспективные направления модернизации Государственной системы научно-технической информации.

А. Ускоренное формирование цифровых информационных ресурсов и их рациональное размещение. Создание распределенных сетевых информационных ресурсов – это наиболее бурно развивающееся направление информатизации научно-промышленной сферы. Развитие коммуникационных возможностей приводит к росту доступной информации, что объективно способствует интенсификации научной деятельности. С учетом активной конвергенции информационных, традиционных библиотечных, компьютерных и телекоммуникационных технологий, цифровые сетевые информационные ресурсы становятся одним из основных источников информации.

В. Создание системы информационных порталов трансфера технологий (по отраслям промышленности). Для развития инновационных процессов в отраслях промышленности исключительно важна информационная поддержка взаимодействия ключевых аудиторий на этапах трансфера технологий инновационного цикла. Насущно необходимо создание проблемно-ориентированного интернет-ресурса, обеспечивающего интерактивное взаимодействие и многофункциональную информационную поддержку участников инновационных процессов. В настоящее время в России реально функционирует только Федеральный портал по научной и инновационной деятельности (www.sci-innov.ru). Его отличает ориентация на весьма ограниченную тематику, определяемую перечнем приори-

тетных направлений развития науки, технологий и техники и перечнем критических технологий РФ.

C. Разработка механизма (технологии) смысловой навигации и поиска знаний в информационных сетях. В настоящее время теория научно-технической информации не располагает методами индустриальной интеграции знаний, представленных в разнородных источниках. При поддержке Российского фонда фундаментальных исследований (проект № 17-07-00153) в ВИНИТИ РАН и БЕН РАН ведутся работы в этом направлении на основе как интеллектуальных методов, так и автоматических методов анализа содержания классификационных систем и их соотношений. Созданы алгоритмы и программный комплекс навигации, поиска и сбора информации на основе связей, зафиксированных в онтологии научного и технического знания [3].

D. Разработка и широкое внедрение технологии интернет-избирательного распространения информации (интернет-ИРИ). Развитие этой технологии как новой системы информационного обслуживания особенно актуально. Она должна базироваться на использовании механизма кластеризации потоков информации из открытых источников и методов построения адаптивных гипермедиа на основе технологии кластеризации неструктурированных данных и обеспечения донесения актуальной, лингвистически обработанной информации до различных целевых групп ее потребления (и отдельных пользователей) в соответствии с их персональными потребностями и ожиданиями. С некоторой долей условности можно говорить о создании ИРИ нового поколения на основе конвергенции телекоммуникационных, компьютерных и информационных технологий. Качественно новый уровень конвергированного ИРИ характеризуется практически неограниченным кругом источников (и пользователей), предельной минимизацией временного лага, высокой целевой избирательностью [4].

E. Разработка САПР информационной поддержки работ инновационного цикла. В современных условиях для создания и производства новой продукции актуально и необходимо использование системы автоматизированного проектирования (САПР) информационного обеспечения работ по всему инновационному циклу (так же, как и конструкторских САПР или САПР технологической подготовки производства). Такая система позволит осуществлять проектирование и эффективное управление комплексным информационным обеспечением во взаимосвязи с изменяющимися задачами и действующими производственными планами по всему распределенному во времени инновационному циклу [1].

F. Создание вебометрической системы цифрового пространства научных библиотек. В качестве основных задач [5, 6], решаемых в процессе создания и функционирования такой системы, выделим:

- повышение роли и значимости публичных и научных библиотек в обществе;
- сохранение и развитие функциональной деятельности библиотек (в зависимости от их типа и вида), поддержание позитивного имиджа в мировом web-пространстве;
- совершенствование (опосредовано) состава и структуры фондов, оптимизация комплектования библиотек;
- интенсификацию процессов цифровизации фондов библиотек;
- стимулирование процессов диверсификации библиотечных услуг и продуктов в цифровой среде; мониторинг и поддержку принятия управленческих решений;
- социологический мониторинг культурного и образовательного предпочтения россиян;
- формирование интегральной оценки уровня и рейтингового распределения библиотек.

G. Разработка и широкое внедрение систем автоматического перевода текстов. В последнее десятилетие стал доминировать статистический подход к машинному переводу. Перевод генерируется на основе статистических моделей, параметры которых являются производными от анализа двуязычных корпусов текста (*text corpora*). Компьютеры оценивают статистические закономерности в больших массивах ранее накопленного цифрового контента. Самообучение компьютера осуществляется посредством анализа достаточно большого (сотни тысяч) количества параллельных текстов – содержащих одинаковую информацию на разных языках. Например, Евросоюз и ООН выпускают множество текстов документов на всех основных языках стран-участниц [7]. Основным преимуществом статистических систем является их свойство не отставать от развития и подвижности языка: если в языке происходят какие-либо изменения, система сразу это распознает и самостоятельно обучается, при этом качественно перевод отличается гладкостью [8, 9].

H. Производство информационно-аналитических продуктов и услуг на основе аналитической постобработки информации с использованием методов наукометрии, анализа данных и компьютерного моделирования. Технологии постобработки могут применяться в исследованиях и разработках, связанных с такими задачами как: а) прогнозирование динамики изменения показателей многомерных технико-экономических объектов и процессов во времени; б) сопоставительный анализ уровня научных исследований, инновационных разработок; в) выявление эмпирических закономерностей, объективно существующих в экономике. Для решения научных и масштабных технико-экономических задач значительные перспективы имеет синтез методов постобработки информации, виртуального моделирования и технологий *Big Data* (Большие Данные), что обеспечит создание качественно новых, на порядки более эффективных, чем раньше, методов аналитической обработки информации, макропроектирования, прогнозирования научно-технических, экономических и социальных процессов, а также комплексной оценки рисков техногенного, природного и социального характера.

I. Значительное ускорение темпов развития широкополосного доступа в Интернет и суперкомрьютинга (основы технологий *Big Data*). Высокоскоростные сети являются базовым элементом разви-

тия распределенных и облачных вычислений, перспективных технологий Больших Данных. По экспертным оценкам рост сети широкополосного доступа (10 Гбит/с) на 10% приводит к увеличению ВВП на 1%, при этом удвоение средней скорости передачи данных в стране увеличивает ВВП на 0,3%. Это свидетельствует о том, что создание широкополосных сетей оказывает непосредственное влияние на развитие национальной информационной инфраструктуры и экономики в целом. Пользователями Интернета в России являются ~82 млн чел. (66% населения), 50 млн из них выходят в Интернет с помощью мобильных устройств (55% смартфоны, 41% планшеты). С высокой степенью вероятности можно прогнозировать рост числа обращений (и сервисов) к базам данных с научно-технической информацией через мобильные устройства [10].

J. Развитие системы подготовки кадров на основе вузовского потенциала и путем поствузовского образования, в частности, систем подготовки специалистов-аналитиков, специалистов по *IT*-технологиям, поиску и мультипликативной обработке научно-технической и технико-экономической информации.

В заключение данного подраздела необходимо отметить, что быстрый рост глобальной сети, количества компьютерных систем, лавинообразное увеличение цифровых данных объективно влечет возрастание рисков и различного рода угроз целостности информации.

Проблемы защиты информации затрагивают различные аспекты ее представления, хранения и обработки, а также вопросы выбора и реализации методов и средств защиты (прежде всего от несанкционированного использования). Система мер защиты информации требует комплексного подхода и включает не только применение технических и программных средств, но и использование организационно-правовых мер защиты. Необходимость обеспечения защищенности и надежности функционирования информационных систем приводит к пониманию целесообразности включения функций защиты в состав основных функций информационных систем ГСНТИ. Информационная безопасность фактически становится одной из характеристик информационных систем.

МАКРОСТРУКТУРА КОМПЛЕКСА РАБОТ ПО МОДЕРНИЗАЦИИ ГСНТИ

С системных позиций кратко рассмотрим структуру и состав комплекса мероприятий и работ [11] по модернизации существующей Государственной системы научно-технической информации.

1. Инвентаризация, аудит, анализ и оценка по направлениям:

- информационная инфраструктура (включая негосударственные системы);
- организационная структура, решаемые задачи, управление, финансирование;
- традиционные и цифровые информационные ресурсы, их соотношение и динамика;
- кадровые ресурсы (численность, качественный уровень, тренды).

2. Сопоставительный анализ, оценка и подготовка пакета предложений и рекомендаций по функциональным направлениям и задачам ГСНТИ на основе специально созданного комплексного структурированного аналитического обзора зарубежных национальных информационных систем. Особый интерес представляют национальные информационные системы Франции и Германии (информационная инфраструктура, ресурсы, финансирование, управление развитием).

3. Разработка характеристической модели существующей и перспективной научно-промышленной сферы: структура, основные задачи, тенденции, приоритеты.

4. Анализ состояния и тенденций развития:

- информационной среды (в том числе научные цифровые ресурсы, автоматически генерируемые данные, СМИ, социальные научные сети);
- информационных и телекоммуникационных технологий (в том числе виртуальное моделирование, мобильные приложения, технологии *Big Data*, широкополосный доступ).

5. Анализ и систематизация факторов-детерминант неэффективного использования информационных ресурсов. Подготовка предложений и рекомендаций по минимизации (нейтрализации) информационных «барьеров».

6. Разработка перспективной концептуальной модели Государственной системы научно-технической информации. Базовые концептуальные положения:

a) смена парадигмы организации информационного обеспечения и функционирования – от иерархической к сетевой;

b) конвергенция информационных, библиотечных, компьютерных и телекоммуникационных технологий. Самоорганизация (в смысле адаптивности структуры и функциональных ролей участников) глобальной сетевой институциональной среды;

c) информационная поддержка взаимодействия ключевых аудиторий при проведении научных исследований на этапах инновационного цикла и трансфера технологий (социальные научные сети и СМИ);

d) автоматизированное проектирование и управление комплексным информационным обеспечением исследований и разработок. Управление знаниями и информационная поддержка принятия решений;

e) углубленная информационно-аналитическая постобработка информации, прогнозирование, компьютерное моделирование.

7. Формирование и систематизация пула актуальных проблемно-ориентированных макрозадач (направлений) информационного обеспечения, в том числе:

- создание эффективных методов и средств управления процессами информационной поддержки цикла исследование – разработка – производство;
- внедрение новых технологий постобработки информации и производство информационно-аналитических продуктов и услуг с использованием методов наукометрии, эконометрии, многомерного анализа данных и компьютерного моделирования.

8. Разработка трехлетней Программы модернизации ГСНТИ, включающей разделы:

- Цели. Задачи. Этапы реализации. Ресурсы

- Состав (и ответственность) организаций-соисполнителей
- «Дорожная карта» реализации Программы
- Система организационно-правовых отношений и взаимодействия государственных и негосударственных структур в сфере научно-технической информации
- Оценка совокупных бюджетных (и небюджетных) затрат по этапам
- Система показателей и индикаторов достижения целей программы.

9. Задачи и мероприятия в рамках информационного взаимодействия стран СНГ (БРИКС, ШОС):

- опорные организации – генераторы баз данных и национальные аналитические центры;
- организационное и нормативно-правовое обеспечение, управление и координация;
- электронные информационные ресурсы;
- опорная телекоммуникационная инфраструктура;
- навигация, поиск, защита информации.

Результаты модернизации Государственной системы научно-технической информации будут иметь решающее значение для развития: 1) фундаментальных и прикладных исследований и разработок; 2) высокотехнологических отраслей промышленности; 3) среднего и высшего профессионального образования; 4) органов власти и управления; 5) международного сотрудничества в информационной сфере.

О СОЗДАНИИ НАЦИОНАЛЬНОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Состояние и перспективы развития информационного рынка в России во многом определяются соответствием сложившейся информационной инфраструктуры современным требованиям, теми предпосылками, которые существовали до начала развития рыночных отношений в стране, а также изменениями в информационной деятельности, которые происходят в последние годы. На мировом рынке информации можно выделить основные секторы, которые характерны также для России [12]:

1. Сектор деловой информации, охватывающий:

- биржевую и финансовую информацию (котировки ценных бумаг, учетные ставки, курсы валют, рынок товаров и капиталов, цены, инвестиции);
- экономическую и статистическую информацию (динамика, тренды, модели, прогнозы, демографическая статистика), деловые новости в области экономики и бизнеса;
- коммерческую информацию (компании, фирмы, направления их работы, их финансовое состояние, персоналии, сделки, продукция).

2. Сектор научно-профессиональной информации: научно-технической, технико-экономической, медицинской, нормативно-правовой и другой информации.

3. Сектор массовой, потребительской информации (расписание транспорта, сведения о погоде, предложения по обмену, покупкам и продажам, товары и услуги, игры, справочники, курсы валют и т.п.).

4. Сектор социально-политической информации, обеспечивающий органы государственной власти и управления статистической, социальной, демографической, юридической, архивной и специальной информацией.

Несмотря на довольно большое число негосударственных информационных организаций, работающих на коммерческой основе, рынок, если исходить из оценки роли организаций в генерации БД с точки зрения новой «оригинальной» информации, создаваемой непосредственно на основе обработки первоисточников, а не слияния и реструктурирования «чужих» массивов, по-прежнему определяется весьма ограниченным числом крупных государственных структур – информационных центров и библиотек. Эти структуры представлены ведущими государственными информационными организациями, системой информационного обеспечения органов государственной власти и управления, сетью библиотек, специализированными институтами и центрами (ВИНИТИ, ИНИОН, ЦИТИС, ВИМИ, ГПНТБ и др.). В совокупности эти организации обеспечивают российское общество не менее чем 90% «оригинальной» достоверной/выверенной информации, тогда как роль многочисленных негосударственных организаций ограничивается ее «переупаковкой» и маркетингом, добавляя в совокупные информационные ресурсы России не более 10% [11]. Малый информационный бизнес не в состоянии удовлетворить информационные потребности экономики и общества самостоятельно, и в основе успеха его работы лежат возможности использования ресурсов, создаваемых в основном рамках государственной системы. Национальная информационная система (НИС) включает, по определению, как государственные структуры, так и негосударственные, коммерческие структуры. Институциональное формирование перспективного концептуального (структурно-организационного) облика НИС, в рамках современных реалий, может осуществляться лишь на основе конвергенции негосударственного сегмента с полноценной развитой Государственной системой научно-технической информации, в полной мере финансируемой из средств федерального бюджета. Рассчитывать на значимую поддержку со стороны крупного бизнеса было бы, по меньшей мере, наивно (как показывает опыт становления НИС в развитых странах). Государственная информационная политика в области информации и телекоммуникаций должна быть сконцентрирована на поддержке действующих институтов (структур) и каналов распространения информации в России. Эту инфраструктуру следует сохранять и развивать, меняя формы и методы работы. Таким образом ГСНТИ интегрируется в новую рыночную среду, что является одновременно основой и гарантией того, что Россия не потеряет накопленного информационного потенциала, не останется без собственных национальных информационных ресурсов при создании современной цифровой экономики.

ЗАДАЧИ ВИНИТИ РАН КАК ВЕДУЩЕЙ ОРГАНИЗАЦИИ ГСНТИ

При создании ВИНИТИ РАН в 1953 г. базовая концепция заключалась в организации национального центра реферирования мирового потока литературы по всем направлениям фундаментальных и прикладных исследований в естественных и технических науках.

Совместным приказом-распоряжением РАН и Миннауки РФ от 14 октября 1998 г. № 192/15 на ВИНИТИ возложены обязанности головной организации ГСНТИ. В соответствии с приказом-постановлением Министерства промышленности, науки и технологий РФ и Президиума РАН от 03 марта 2004 г. № 73/25 функциональные задачи ВИНИТИ в области информационного обеспечения научно-промышленной сферы и координации работ по созданию и развитию общесистемной нормативно-методической базы ГСНТИ были подтверждены и дополнены следующими позициями [12]:

- генерация и развитие политематического банка данных по естественным и техническим наукам как составной части государственных информационных ресурсов;
- научно-информационное и аналитическое обеспечение научных исследований по естественным и техническим наукам, а также в области национальной экономики и образования в соответствии с федеральными и региональными программами и проектами;
- разработка научно-методологических основ информатизации общества и инновационной деятельности, направленной на обеспечение социально-экономического развития и национальной безопасности Российской Федерации;
- создание: а) концептуальных основ и методологических подходов к оценке эффективности процессов информатизации общества; б) программных средств построения интеллектуальных информационных систем для поддержки научной, производственной и образовательной деятельности;
- ведение и издание Государственного рубрикатора научно-технической информации (ГРНТИ) и банка эталонных таблиц Универсальной десятичной классификации (УДК) на русском языке;
- организация мониторинга информационной продукции и услуг органов НТИ, ведение и издание сводного каталога органов НТИ России и стран СНГ, приобретение и использование зарубежной научно-технической литературы организациями, входящими в ГСНТИ.

Постобработка больших массивов научно-технической и технико-экономической информации с использованием статистических методов, методов анализа данных, позволяет на основе политематического банка данных ВИНИТИ выявлять статистические закономерности, выражающие зависимости между распределениями различных параметров исследуемых систем и процессов и характер изменения распределений во времени [13].

Политематический БнД ВИНИТИ содержит свыше 36 млн записей (глубиной ретроспективы по некоторым предметным областям до 15 лет). Использование статистических методов при аналитической постобработке реферативной и библиографической информации такого объема представляется весьма перспективным для решения ряда задач, в числе которых:

- анализ структуры отечественной и мировой науки;
- определение тенденций и процессов, происходящих в мировой и региональной науке;
- выявление наиболее актуальных или, напротив, теряющих свою актуальность научных направлений;
- отслеживание генезиса конкретных научных идей и истории их развития;
- определение продуктивности работы исследователей в конкретной научной области и эффективности материальных затрат в этой области;
- анализ структуры научного сообщества и науки как социального организма.

Помимо этого ВИНИТИ обрабатывает информацию по химическим структурам. Банк данных структурной химической информации содержит 646 тыс. химических соединений и более 150 тыс. химических реакций. Это очень ценный, если не сказать, уникальный ресурс.

Однако следует констатировать, что кризисные явления в российской науке постсоветского периода не обошли стороной и ВИНИТИ. За период с конца 1980-х по 2018 г. наполнение Реферативного журнала (РЖ) упало с 1 млн 400 тыс. документов в год до 638 тыс. Из-за систематического недофинансирования значительно сократилась численность квалифицированных референтов, переводчиков, редакторов. Качество любого РЖ зависит от таких факторов как оперативность отражения публикаций, полнота охвата заявленной тематики и основных изданий и, особенно, полнота составляемых рефератов, степень разработанности справочного и поискового аппарата, глубина и адекватность рубрицирования. Несколько утрируя, можно сказать, что с начала 1990-х гг. ВИНИТИ занимается только переработкой информации, а никак не ее распространением. Интерфейс пользователя Банка данных ВИНИТИ не соответствует современным требованиям. Как негативный фактор следует отметить также фактическую утрату контроля со стороны регулятора (со стороны государства – это Президиум РАН) за деятельностью ВИНИТИ в сфере как материально-финансового обеспечения, так и целеполагания. Здесь следует отметить, что еще в советский период основным потребителями информационной продукции и услуг ВИНИТИ были отраслевые НИИ, КБ, проектные организации, промышленность, вузы – это примерно 93%, доля Академии наук не превышала 5-7%. Кроме того необходимо принимать во внимание, что даже в 2018 г. российская фундаментальная наука (с учетом инфляции) не вышла по финансированию на уровень 2014 г. А по доле ВВП, которую Россия расходует на фундаментальные исследования, мы находимся на уровне Мексики и ЮАР [14].

Предпринятые руководством страны шаги по реформированию Российской академии наук, развитию науки и промышленности актуализировали проблему структурно-функциональной модернизации ВИНИТИ и совершенствования его информационной деятельности в соответствии с новыми вызовами и задачами создания инновационной цифровой экономики России. Задача реферирования мирового потока научной литературы не утратила своего значения, но сущест-

венно изменилась. Следует отметить, что в развитых странах по-прежнему издается около 3 тыс. реферативных журналов, они выходят в электронном виде и выполняют информационно-поисковые и науковедческие функции, а рефераты в них становятся в основном индикативными. Это нейтрализует действие закона Брэдфорда о рассеянии публикаций определенной тематики по всему массиву журналов, способствует развитию национальной науки, выработке собственной терминологии и собственной информационной политики.

В части научного сообщества сложившаяся ситуация формирует дискурс: ***нужен ли вообще ВИНИТИ и, в частности, Реферативный журнал?*** [15]. Как правило, ученые активно ищут необходимую информацию в Интернете по 5−10 основным журналам. Считается, что это занимает немного времени. Например, в каждом выпуске РЖ ВИНИТИ «Акустика» количество отражаемых источников приближается к 100. И чтобы быть в курсе развития этой науки, потребуется уже немало времени для поиска информации в Интернете. В РЖ же она рассортирована по рубрикам и снабжена поисковым аппаратом. Это резко снижает затраты времени. Научно-техническая информация вообще дисперсна по источникам и неоднородна по полноте и качеству. Следствием этого основными видами научных коммуникаций стали систематические контакты исследователя с ограниченным кругом узких специалистов и изданий. Такая практика приводит к тому, что за пределами поля зрения ученых остаются даже смежные научные области, но при этом создается иллюзия знания всего необходимого, что не может не беспокоить, и беспокойство по этому поводу уже открыто высказывается в развитых странах, так как снижение уровня и качества информационной поддержки ученых ведет к снижению результативности научного труда, ставит перед информатикой проблемы навигации по накопленным информационным ресурсам и их обработки. По существу речь идет об углублении интеллектуальной аналитической переработки информационных потоков, создании принципиально новых информационно-аналитических продуктов и эффективных средств навигации с использованием нетрадиционных лингвистических и программно-технологических средств.

Есть один немаловажный вопрос: ***Нужен ли наш РЖ мировому научному сообществу?*** Без сомнения, нужен, если обратить внимание на полноту отражения в нем русскоязычной научной литературы. Именно это и хотят получать зарубежные ученые, и на это следует ориентироваться в отражении наших даже малотиражных изданий.

Так что же должен делать ВИНИТИ РАН? Новая концептуальная основа развития информационной деятельности Института достаточно детально представлена в работе [16]. Здесь сформулируем лишь некоторые наиболее важные и актуальные направления информационно-технологической модернизации ВИНИТИ.

- Реализация режима открытого бесплатного доступа из Интернета к Банку данных ВИНИТИ (для российских научных и образовательных организаций). Мировой опыт показывает, что наиболее быстро развиваются те ресурсы, доступ к которым бесплатен. Это послужит стимулом для улучшения качества предлагаемого информационного продукта.
- Создание центрального сервера ГСНТИ с размещенной на нем навигационной системой по информационным ресурсам отечественных (и зарубежных) информационных центров, научных библиотек, порталов научных организаций.
- Обеспечение полной реферативной переработки всей русскоязычной научно-технической литературы (НТЛ). Реализация версии БнД по русскоязычной НТЛ для зарубежных пользователей (на платной основе) с использованием современных технологий машинного перевода и биллинговой системы расчетов.
- Разработка и внедрение системы поддержки принятия решений (СППР) по бюджетному финансированию тематических направлений исследований РАН (~8000) с использованием критериев и методов наукометрии и анализа данных (с учетом приоритетности и научно-технического потенциала научных организаций) [17, 18].
- Создание и ведение БнД формализованных характеристик научных организаций Российской академии наук – для целей управления, оптимизации процессов финансирования исследований, научно-технического прогнозирования, развития экспертной деятельности, мониторинга текущего состояния.
- Реализация режимов информационного обслуживания на базе: а) электронного РЖ (более информативного, с улучшенной визуализацией); б) сетевого избирательного распространения информации нового поколения [16].

В заключение данного подраздела следует отметить, что ВИНИТИ РАН является базовой организацией государств – участников СНГ по межгосударственному обмену научно-технической информацией (Решением Совета глав правительств Содружества независимых государств (СНГ) от 19 ноября 2010 г.)

В рамках СНГ межгосударственный обмен научно-технической информацией предусматривает:

- совместное формирование и использование информационного ресурса, обмен национальными ИР, которые содержат сведения об объектах интеллектуальной собственности, результатах НИОКР, инновационной деятельности;
- выполнение совместных научно-технических программ, проектов межгосударственного сотрудничества в сфере НТИ, включая подготовку и переподготовку кадров в этой сфере;
- создание сводных электронных каталогов информационной продукции и услуг национальных информационных центров, а также формирование интегрированной системы доступа к информационным ресурсам стран-участниц.

Необходимо подчеркнуть, что в настоящее время ВИНИТИ РАН является единственным крупным информационным центром в России, информационные потоки которого направлены вовнутрь, а не во вне страны.

ЗАКЛЮЧЕНИЕ

1. Информация – это ресурс, играющий доминирующую роль в системе глобального мирового экономического и социального развития. Информационные ресурсы России являются стратегическими ресурсами, аналогичными по значимости материальным, сырьевым, энергетическим, трудовым и финансовым. Однако на настоящем этапе мы имеем: оскудение ресурсов библиотек; постепенное свертывание производства национальных информационных ресурсов крупнейшими информационными центрами страны; дублирование выделяемых скромных средств на библиотечно-информационную деятельность, недостаточно высокие темпы развития информационной инфраструктуры, телекоммуникационных сетей и технологий. Именно поэтому назрела модернизация системы информационного обеспечения научно-промышленной сферы.

2. Государственная система научно-технической информации, даже в своем нынешнем состоянии, способна и готова расширить свое участие в обеспечении научных исследований, инновационной и образовательной деятельности, а также других социально значимых задач Цифровой экономики и перейти на новый качественный уровень информационной поддержки отечественной науки и промышленности. В России еще имеется необходимый научный и технический потенциал для формирования соответствующей требованиям времени информационной инфраструктуры, создания на базе ГСНТИ полнофункциональной Национальной информационной системы, и этот потенциал необходимо реализовать в кратчайшие сроки.

3. Широкое применение в структуре ГСНТИ цифровых информационных ресурсов, новых информационных технологий содействует более эффективному решению задач информационного обеспечения инновационной деятельности. Информационный компонент научно-технического комплекса России прямо или косвенно отражается в проявлении эффекта:

- мультипликации использования новых научно-технических результатов, знаний и информационных ресурсов;
- комплексного подхода к инвестициям и инновациям в научно-промышленной сфере;
- экономии общественно необходимого времени и материально-технических ресурсов за счет типовых проектных решений;
- трансфера технологий и использования частных технических решений (в разных отраслях).

4. Для реализации рассмотренных нами направлений развития информационной деятельности ГСНТИ имеются значительные заделы и достаточный научно-технический потенциал. При этом, безусловно, необходимы безотлагательные шаги по формированию новой высокопроизводительной информационно-телекоммуникационной инфраструктуры, укреплению и обновлению кадровых и технических ресурсов. Очевидно, что для реализации этих масштабных задач необходимо привлечение дополнительных средств из госбюджета, государственных и целевых федеральных программ, научных и технологических фондов. Представляется целесообразным сформировать для ВИНИТИ, как головной организации ГСНТИ, целевое госзадание на общую координацию работ, разработку «дорожной карты», развитие общесистемной нормативно-методической базы, мониторинг формирования и использования цифровых информационных ресурсов, проведение научных исследований проблем сбора, аналитико-синтетической переработки, хранения, поиска, распространения и использования научно-технической информации. Необходимой предпосылкой успешного решения представленного комплекса задач является концентрация полномочий и ответственности по модернизации национальной информационной инфраструктуры в рамках одного федерального ведомства.

5. Помимо информационной поддержки технологического развития организаций и предприятий промышленности, информационная деятельность уже в современном ее состоянии способна на большее. Прежде всего, в области моделирования различных траекторий социально-экономического развития, перехода на новые пакеты технологий и технологические платформы и оценки возможных результатов управленческих решений, а также связанных с этим масштабных текущих и капитальных затрат [19].

6. Защита информационного пространства России является одним из приоритетных направлений обеспечения национальной безопасности и Национальной программы «Цифровая экономика РФ». Организация национального сегмента сети Интернет необходима, она экономически и стратегически обоснована [20]. В России необходимо создать: а) свою ключевую инфраструктуру Интернета, включая национальные корневые сервера, национальную систему маршрутно-адресной информации; б) свои электронные компоненты, телекоммуникационное оборудование, системное программное обеспечение.

СПИСОК ЛИТЕРАТУРЫ

1. Сюнтюренко О.В., Булычева О.С., Концептуальный облик перспективного технологического пакета информационной поддержки наукоемкого производства // Научно-техническая информация. Сер. 2. – 2016. – № 4. – С. 1-10.
2. Короткевич Л.С. Государственная система научной и технической информации в СССР: итоги и уроки. – М.: ВИНИТИ. – 1999. – 273 с.
3. Антошкова О. А., Белоозеров В. Н., Дмитриева Е. Ю. и др. Построение онтологии информационных ресурсов в виде сети библиографических классификаций // Перспективные направления исследований и критические технологии в классификационных системах: материалы научно-практической конференции с иностранным участием (Москва 25-27 окт. 2017 г.). – М.: ВИНИТИ, 2017. – С. 20-25.
4. Сюнтюренко О.В. Перспективы использования интернет-СМИ, журналов открытого доступа и социальных медиа в научно-технической сфере // Научно-техническая информация. Сер. 1. – 2015. – № 6. – С. 30-36; Syuntyurenko O.V. Prospects for Using Online Media? Open-Access Journals?

And Social Media Networks in the Field of Science and Technology // Scientific and Technical Information Processing. – 2015. – Vol. 42, № 2. – P. 112-118.
5. Антопольский А.Б. О целесообразности российского национального вебометрического индекса // Научно-техническая информация. Сер. 1. – 2014. – № 2. – С. 14-18.
6. Булычева О.С., Сюнтюренко О.В. Концептуальные положения и предпосылки создания вебометрической системы цифрового пространства библиотек // Сборник Президентской библиотеки. Сер. «Электронная библиотека». – 2018. – Вып. 8. – С. 19-31.5.
7. Brynjolfsson E., McAfee A. The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. – New York: Norton & Company, 2016. – 320 p.
8. Дроздова К.А. Машинный перевод: история, классификация, методы // Материалы III Междунар. науч. конф. «Филологические науки в России и за рубежом» (Санкт-Петербург, июль 2015 г.). – СПб: Свое издательство, 2015. – С. 139-141. – URL https://moluch.ru/conf/phil/archive/138/8497/ (дата обращения: 28.12.2018).
9. Колганов Д.С., Данилов Е.А. Обзор аналитической, статистической и нейронной технологии машинного перевода // Международный студенческий научный вестник. – 2018. – № 3-2. – URL: http://eduherald.ru/ru/article/view?id=18262 (дата обращения: 28.12.2018).
10. Сюнтюренко О.В. Факторы-детерминанты неэффективного использования информационных ресурсов в научно-технической деятельности // Научно-техническая информация. Сер.1. – 2017. – № 7. – С. 1-12; Syuntyurenko O.V. Determinants of the Ineffective Use of Information Resources in Scientific and Technological Activities // Scientific and Technical Information Processing. – 2017. – Vol. 44, № 3. – P. 159-169.
11. Сюнтюренко О.В., Каленов Н.Е., Цветкова В.А. Актуальные задачи модернизации системы информационного обеспечения научно-промышленной сферы // Информация и инновации. – 2018. – Т. 13, № 2. – С. 7-17.
12. Арский Ю.М. Земля и ее инфосфера. – М.: ВИНИТИ РАН, 2011. – 356 с.
13. Борисова Л.Ф., Сюнтюренко О.В. Реферативный банк данных ВИНИТИ РАН: перспективы постобработки информации с использованием методов анализа данных // Научно-техническая информация. Сер. 1. – 2007. – № 11. – С. 6-11; Borisova L.F., Syuntyurenko O.V. VINITI RAN Abstract Database: Prospects of Information Postprocessing Using Methods of Data Analysis // Scientific and Technical Information Processing. – 2007. – Vol.34, № 6. – P. 278-283.
14. Прошло пять лет // Научное сообщество. – 2018. – № 6-7(202-203). – С. 15.
15. Шамаев В.Г. Реферативный журнал ВИНИТИ РАН и проблемы информационного обеспечения российской науки // Троицкий вариант. – 2011. – № 87(13 сентября). – С.10. – URL: http://trv-science.ru/2011/09/13/referativnyj-zhurnal-viniti-ran (дата обращения 04.02.2019).
16. Биктимиров М.Р., Гиляревский Р.С., Сюнтюренко О.В. Новая концептуальная основа развития информационной деятельности ВИНИТИ РАН // Научно-техническая информация. Сер. 1. – 2016. – № 1. – С. 1-8; Biktimirov M.R., Gilyarevskii R.S., Syuntyurenko O.V. A New Conceptual Basis for the Development of the Information Activities of the All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences // Scientific and Technical Information Processing. – 2016. – Vol. 43, № 1. – P. 1-7.
17. Сюнтюренко О.В., Гиляревский Р.С. Использование методов наукометрии и сопоставительного анализа данных для управления научными исследованиями по тематическим направлениям // Научно-техническая информация. Сер. 2. – 2016. – № 12. – С. 1-12.
18. Сюнтюренко О.В. Финансирование фундаментальных исследований: концептуальный облик системы поддержки принятия решений с использованием методов наукометрии и анализа данных // Информатика и ее применения. – 2018.– Том 1, Вып.1. – С. 118-126.
19. Родионов И.И., Гиляревский Р.С., Цветкова В.А. Информационная деятельность как инфраструктура национальной экономики. – СПб: Алетея, 2016. – 223 с.
20. Черешкин Д.С., Смолян Г.Л. Информационная инфраструктура и информационная безопасность (Эволюция представления о предметной области за последние 50 лет) // Материалы международной научно-практической конференции «Информационная безопасность: вчера, сегодня, завтра» (Москва, 12 апреля 2018 г.). – М., 2018. – С. 67-73.

Материал поступил в редакцию 15.05.19.

Сведения об авторах

СЮНТЮРЕНКО Олег Васильевич – доктор технических наук, профессор, ведущий научный сотрудник ВИНИТИ РАН
e-mail: olegasu@mail.ru

ДМИТРИЕВА Елена Юрьевна – кандидат технических наук, заведующая научно-методологическим отделением ВИНИТИ РАН,
e-mail: niipio@mail.ru

ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ

УДК 004.65:002.6

А.О. Еркимбаев, В.Ю. Зицерман, Г.А. Кобзев, А.В. Косинов

Курирование цифровых научных данных

Изучена роль процесса курирования в поддержке хранилищ научных данных. Показано, что в дисциплинах с интенсивным использованием данных (науки о Земле, биология, материаловедение и т. п.) курирование является существенным элементом научной работы. Масштаб и значимость курирования для научных архивов и баз данных обусловили появление обширной литературы и стандартов, регламентирующих требуемые действия. Рассмотрены в деталях меры по сохранности данных, их очистке от искажений, оценке качества и детализированному описанию. Изучены достоинства и недостатки действующих стандартов научных метаданных. Показано, что постоянное соблюдение всех требований, регламентирующих процесс курирования, способно обеспечить не только сохранность, но и непрерывное обогащение ценности научных данных.

Ключевые слова: *интенсивное использование данных, четвертая парадигма, база данных, репозитарий, качество данных, курирование данных, метаданные*

ВВЕДЕНИЕ

Согласно словарю синонимов[1] термин «курирование» означает «присматривание, наблюдение, опекание», тогда возникшее по аналогии понятие «цифровое курирование» должно означать такую же «заботу» в отношении хранилища цифровых данных. Однако, уже на уровне дефиниции проявляется специфика цифрового материала: как Википедия (https://en.wikipedia.org/wiki/Digital_curation), так и Центр цифрового курирования (*Digital Curation Centre* – DCC, www.dcc.ac.uk) расширяют трактовку понятия условием «добавления ценности» к исходным данным. Это сразу же свидетельствует о том, что цифровое курирование не сводится к текущему надзору за состоянием фондов, а включает достаточно серьезную аналитическую работу. Её значимость и масштаб подтверждают создание специализированных центров типа DCC, международного журнала «*International Journal of Digital Curation*» (www.ijdc.net), издание многочисленных руководств и справочников [1-4], а также действующие стандарты и т.д.

Потребность в цифровом курировании существует везде, где собраны большие коллекции, например, оцифрованных фильмов или книг. Но особые требования возникают в дисциплинах, получивших название "*data-intensive science*" или "*e-Science*", например, в науках о Земле или биологии [5, 6]. Необходимость их выделения в особую категорию впервые высказана Дж. Греем в 2007 г.[2], предложившим концепцию «четвертой парадигмы», дополняющей три традиционные: эксперимент (или наблюдение), теория, моделирование. При крайнем различии предметных областей (например, астрономии и биохимии) возникло методическое единство – необходимость формировать и хранить обширные массивы данных с их последующим использованием для анализа и моделирования. Как следствие, процедуры курирования вошли составным элементом в практику любой из *e-Science*. Избегая точных дефиниций, можно сказать, что курирование преследует три основных цели: сохранность данных, их непрерывную очистку от искаженных (неполных или несогласованных) данных и, наконец, описание данных, т. е. назначение и развитие системы метаданных.

Возникает вопрос, почему обычное «поддержание порядка» в данных привлекло столь обширное внимание специалистов как в информатике, так и в предметных областях. Прежде всего, хранение цифрового контента опирается на программно-аппаратное обеспечение и множество форматов данных, подверженных достаточно быстрому моральному старению за счет несовместимости с современными платформами, прекращения методи-

[1] Словарь синонимов русского языка – онлайн подбор. – URL: https://sinonim.org/.

[2] Доклад Дж. Грея «A Transformed Scientific Method» приведен в [6].

ческой поддержки и т.п. На протяжении одного поколения такие средства как перфорированная лента, дискеты, CD-ROM, DVD быстро сменяли друг друга, затрудняя или исключая доступ к вышедшим из употребления. Препятствовать быстрому устареванию можно за счет непрерывного управления архивным фондом со сменой среды хранения, форматов и программ. Для сравнения заметим, что практика хранения древних рукописей или музейных экспонатов обеспечивает их сохранность на протяжении многих веков при минимальных и вполне доступных условиях.

Второй момент – нарастающие объемы цифровых архивов, используемых в науке, что и легло в основу концепции «четвертой парадигмы». Как следствие – методы курирования, традиционные в практике библиотек, архивов или музеев, оказались слабо пригодны при таких масштабах, которые требуют создания специальных средств, технологий и стандартов, составляющих основу курирования на современном уровне.

И наконец, сам характер научных данных предъявляет требования, теряющие смысл в отношении цифровых данных, порождаемых бизнесом или массмедиа. Суть дела в том, что все чаще публикация научных данных происходит вне традиционного издательского процесса, заполняя многочисленные базы данных (БД) и репозитарии. В связи с этим авторы монографии [6] отмечали, что «данные и программы должны стать неотъемлемой частью архива науки – объектами первого уровня, которые тоже требуют систематического управления и курирования… Эта тенденция отражается в акценте на курировании и многократном использовании данных в различных инфраструктурах и программах e-Science». По существу, журнальные публикации вмещают лишь аннотирование материала, в то время как многочисленные файлы с данными и их визуализацией находятся в репозитариях с устойчивым идентификатором (ID), допускающим независимое цитирование. Задача курирования – компенсировать недостатки изолированной публикации данных, когда они теряют контекст, разъясняющий их генезис, достоверность, единицы измерения, обозначения и проч. Эта часть работ по курированию может рассматриваться как экспертное описание наборов данных, что заметно выходит за рамки старой деятельности библиотек или архивов. Более того, курирование как средство обеспечения науки предполагает за куратором необходимость обращаться к созданию (или использованию) индексирующих систем, онтологий, стандартов метаданных и множества других средств, обеспечивающих согласованность в работе исследовательских групп и репозитариев [5].

К ИСТОРИИ ПОНЯТИЯ «КУРИРОВАНИЕ ДАННЫХ» (*data curation*)

Идея о курировании данных как деятельности особого рода, по-видимому, впервые прозвучала в 1993-м г. в докладе Министерства энергетики США, посвященном роли БД в проекте генома человека [7], в котором содержалось предложение организовать подготовку научных работников по новой специальности под названием «Курирование и контроль качества данных». Однако это подчеркивало ограниченное понимание курирования только как контроля качества информационного ресурса. Заметно более широкий взгляд на курирование данных высказан в статье Дианы Зорич [8], отметившей глубокое единство в задачах курирования библиотек, музеев и цифровых архивов. Со ссылкой на доклад Министерства энергетики, она предложила считать задачей куратора перманентную проверку данных на согласованность, долгосрочное качество и актуальность, обратив особое внимание на поддержку и обновление тезаурусов, словарей и других средств документации. Главным в ее концепции был перенос акцента на управление вторичными данными, полученными при описании первичных, т. е. на *метаданные*. Фактически с этой публикации курирование признается как особый вид деятельности, применимый к цифровым данным в различных дисциплинах.

Дальнейшая история включения этой концепции в практику управления цифровыми хранилищами связана с семинаром *Digital Curation: digital archives, libraries and e-science seminar*, проведенным в Лондоне 19 октября 2001 г. [9]. Детализацию задач, решаемых в ходе «курирования данных», участники семинара связали с активным заимствованием информационных технологий в различных подходах, издавна реализуемых работниками архивов, библиотек и музеев.

Многолетний опыт хранения физических объектов в библиотеках и музеях открыл для кураторов цифровых коллекций важнейшее требование по непременному обогащению их ценности по мере хранения. Этот опыт подсказывал, что источником новой ценности может служить сопровождающая документация, расширение контекста за счет ссылок на новые ресурсы сети, предметно-ориентированная экспертиза и навыки куратора. Применительно к цифровым коллекциям, это позволило отделить понятие «курирование» от таких технически ориентированных понятий как «архивирование» (*archiving*) и «сохранение» (*preservation*) [10].

Длительное обсуждение в экспертной среде задач курирования позволило придти к относительно согласованному определению этого вида деятельности. Вот как, например, звучит одно из наиболее компактных определений, данных еще в 2003 г. P. Lord и A. Macdonald [10]: *«Деятельность по использованию данных с момента их создания гарантирующих их пригодность для современных целей, обнаружения и повторного использования. Для динамических наборов данных это может означать также непрерывное расширение или обновление, чтобы поддерживать их в соответствии с исходным назначением. Более высокий уровень курирования включает также создание и поддержку связей с аннотацией и другими опубликованными материалами».*

ОСНОВНЫЕ СОСТАВЛЯЮЩИЕ ПРОЦЕССА КУРИРОВАНИЯ

В представленной Центром DCC краткой статье [11] раскрываются решаемые в процессе курирования задачи:

- сохранение и защита данных от потерь и морального износа оборудования и форматов, что важно для невоспроизводимых и особенно ценных данных;

- сохранение доступа к данным после прекращения финансирования или организационной перестройки;
- повторное использование данных;
- информирование о происхождении данных и контекстная поддержка;
- обеспечение доступности и понятности для пользователей средств миграции данных в новые форматы.

Детальный анализ, проведенный в работе [3], позволил выявить до 50 технических и организационных мер, которые составляют курирование цифровых данных; по существу их можно отнести к одному из трех ключевых направлений: обеспечение сохранности данных, очистка, т. е. коррекция искаженных (неполных или несогласованных) данных, и наконец, описание данных, т. е. назначение и развитие системы метаданных. Причем, на каждой из стадий курирования научных данных преследуется не просто «поддержание порядка», но и непрерывное их обогащение за счет привлечения новых программных средств и форматов, а также нового контекста за счет расширения метаданных и/или гиперссылок на новые ресурсы.

Сохранность данных. В табл. 1 приведены некоторые из мер, что были рекомендованы в работе [3] для целей долгосрочного хранения цифровых данных.

В основном, эти меры сводятся к контролю физического состояния среды хранения и преодолению технологического устаревания, т. е. морального износа оборудования и программных средств. Как и физические артефакты (музейные экспонаты или рукописи), цифровой материал тоже подвержен эрозии в виде так называемого *гниения или распада битов* (*bitrot*), частота которого заметно возрастает из-за износа, загрязнений, теплового или радиационного воздействия. Основная мера борьбы с деградацией файлов на уровне битов состоит в периодическом просмотре и регулярном резервировании архива.

Вторая проблема сохранности цифрового архива состоит в преодолении морального износа оборудования, программ и, прежде всего, файловых форматов. При физической сохранности моральный износ означает прекращение методической и программной поддержки прежних версий, неполную совместимость с современными платформами. Основная мера преодоления технологического отставания – миграция форматов (см. табл. 1), т. е. их конверсия в более «свежие» форматы, совместимые с новыми платформами. Помимо устаревания формата, потребность в подобной конверсии может быть связана с желательным отказом от так называемых *патентованных* (*proprietary*) форматов, т. е. связанных с узкой сферой применения, например, в геоинформационных системах или в рентгенографических исследованиях.

В более общем случае миграция сохраняет содержание цифровых объектов в условиях постоянно меняющейся технологии, включая программное и аппаратное обеспечение. При этом миграция отличается от обновления носителей тем, что не полностью воспроизводит всегда точную цифровую копию и оригинальные функции цифрового объекта [4]. Среди других мер поддержки архива в условиях меняющейся технологии предлагается сохранение старых версий платформы, а также *эмуляция*, т. е. имитация устаревших систем на компьютерах будущих поколений[3].

Наряду с мерами по преодолению физического и морального износа системы программно-аппаратной платформы, существуют определенные возможности по отказу от избыточного хранения. Само по себе сокращение объема цифровых данных может рассматриваться как средство их сохранности, поскольку вероятность *bitrot* нарастает с ростом объема файлов.

Таблица 1

Основные меры сохранности цифровых данных

Безопасность среды хранения	Хранение файлов в хорошо сконфигурированной (с точки зрения аппаратного и программного обеспечения) среде хранения, физически защищенной и регулярно резервируемой
Аудит файлов	Контроль и предупреждение возможности цифровой эрозии или повреждений аппаратуры
Инвентаризация файлов	Периодический контроль количества, типов и размеров файлов с протоколированием данных об отсутствующих, дублированных или недоступных файлах
Регистрация программного обеспечения	Сохранение копий устаревших версий программного обеспечения для доступа к данным при изменении платформы
Миграция форматов	Контроль за устареванием форматов файлов и, при необходимости, преобразование устаревших форматов файлов в новые форматы
Трансформация форматов	Преобразование файлов патентованных форматов в открытые форматы, которые расширяют возможности для длительного использования
Эмуляция	Поддержка устаревшей конфигурации системы хранения на современном оборудовании

[3] На практике в цифровых архивах применяются также средства восстановления поврежденной среды, известные под названием *digital archaeology* [4]. Достаточно высокая стоимость, однако, делает их применение оправданным только для особо ценных материалов.

Согласно наблюдению авторов концепции e-Science [12], в научных данных можно почти всегда выделить две категории: эфемерные и стабильные. Первые не могут быть реконструированы и имеют ценность именно в том виде, как были записаны (примеры, приведенные в [12]: осадки, солнечные пятна, плотность озона или цена на нефть). Стабильные данные, напротив, могут быть реконструированы, если есть процедура их получения из некоторых исходных. В частности, можно отказаться от хранения массивов данных, полученных в результате компьютерного моделирования. При сохранении основных метаданных и параметров модели моделирование может быть запущено снова по запросу пользователя. Каждое десятилетие вычисления становятся дешевле на три порядка, что делает замену долгосрочного хранения на повторный счет более выгодным. В работе [5] в качестве примера избыточных данных были приведены обширные таблицы термодинамических свойств, которые рассчитываются по известному уравнению состояния. Другой пример, иллюстрирующий экономию на объеме хранения, – молекулярно-динамические эксперименты, итогом которых являются обширные протоколы, отражающие эволюцию координат и импульсов для большого ансамбля частиц. Сохранив сведения о параметрах потенциала, а также все программные инструменты для моделирования, можно рассчитывать на воспроизводимость результатов тех же численных экспериментов в неопределенном будущем, сняв нагрузку, связанную с хранением обширных данных моделирования.

Очистка данных. Важнейший из элементов курирования, направленный на выявление и исправление разнообразных дефектов в данных, очистка (*cleaning or cleansing*) отличается от более узкого понятия *purging*, предусматривающего лишь освобождение пространства от ошибочных данных для ввода новых. В общем случае процесс под названием *data cleaning*, избавляя данные от синтаксических ошибок, опечаток и пропусков, имеет целью провести коррекцию, обеспечив наибольшую точность в данных. Дефекты в этом случае могут быть вызваны ошибками ввода и дублирования, физическим старением записей, неправильным распределением данных по полям. Искаженные данные резко снижают их качество, подтверждая известную поговорку «мусор на входе, мусор на выходе». Для обработки обширных числовых архивов возможно применение формальных подходов (математическая статистика и методы *data mining*) с минимальным привлечением манипуляций, контролируемых человеком [13–16].

Согласно руководству [15], можно выделить несколько характерных видов искажений: нерелевантность, синтаксические ошибки, пропуски, дубликаты, ошибки в типе данных.

Под нерелевантностью понимают явное несоответствие данных контексту, предусмотренному предметной областью. Простейший пример: наличие одного числа, вместо предполагаемого по смыслу набора данных.

Синтаксические ошибки более формальны. К ним относятся пустые пространства в начале и в конце строки, или строки, не дополненные пробелами или другими символами до определенной ширины. Под опечатками понимают обычные нарушения грамматики в текстовой строке, использование символов или аббревиатур, не соответствующих принятому стандарту. Хороший пример дает обязательное использование прописной латинской буквы *B* для второго вириального коэффициента (распространенная термодинамическая характеристика реальных газов), что означает ошибку при использовании строчной буквы *b*.

Пропуски – наиболее частый вид ошибок, способный привести к сильным искажениям при анализе данных. Авторы руководства по методам очистки данных [15] указывают на два альтернативных подхода при обнаруженных пропусках. Первый предполагает исключение из набора незаполненных позиций, если их число невелико. В противном случае и при случайном расположении пропусков можно удалить весь набор. Второй подход состоит в том, чтобы восстановить пропущенную величину, приписав ей некоторое значение. Наиболее обоснованными выглядят соображения, основанные на статистике, специальный раздел которой [16] посвящен восстановлению отсутствующих данных оценочными значениями, основанными на другой доступной информации (*вменение* или *imputation*). Простейший из рецептов, предложенных в статистике, использование среднего или медианного значения. При большом числе пропусков это не гарантирует отсутствие искажений, хотя оценка медианного значения более устойчива к выбросам. Другой способ восстановления данных, использующий линейную регрессию, также достаточно чувствителен к сильным выбросам. Наконец, в статистике издавна применялся прием под названием «горячая колода» (*hot-deck*), когда отсутствующее значение заменяется из случайно выбранной аналогичной записи, например, на значение непосредственно перед отсутствующими данными. Альтернативой этому служит так называемый метод *cold-deck* с выбором замещающих значений из других наборов данных. После замены пропущенных значений, новый набор данных доступен для стандартного анализа, хотя большое число пропусков все-таки привносит отклонения в результаты анализа. К этой же проблеме относятся и большие выбросы, т. е. значения на краях интервала допустимых значений. Если выбросы связаны с ошибками измерений или ввода, то их следует исключать из выборки, заменяя на другие значения, как при пропусках. Однако даже если выбросы являются частью распределения, их рекомендуют исключать из выборки, чтобы обеспечить более устойчивый результат при использовании статистических методов анализа.

Две другие ошибки – дубликаты и неправильное отнесение типа данных. Дубликаты появляются в наборах данных, за счет как ошибок ввода, так и за счет суммирования данных из нескольких независимых источников [13, 14]. Выявление и удаление дублирующих записей исключает ложное увеличение их статистического веса.

Согласно [15], неправильное отнесение типа данных имеет место, когда численные или категориаль-

ные значения не удается преобразовать к числовому типу данных. В этом случае вместо указанного значения возникает NA[4] с указанием на необходимость исправления. В основном, указанные в данных искажения могут быть выявлены и частично скорректированы, основываясь на формальных признаках.

Аттестация качества данных. Результат очистки данных проверяется, как правило, по тому, насколько очищенные данные удовлетворяют критериям качества (см., например: https://en.wikipedia.org/wiki/Data_cleansing). Поэтому эффективное курирование не может обходиться без априорного назначения некоторого набора критериев качества, включаемых в набор метаданных. Их выбор, зависящий от предметной области и принятой стратегии курирования, в деталях рассмотрен в нашей работе [5]. Применительно к научным данным, наилучшая оценка качества сочетает указание неопределенности (или ошибки) с проведением сертификации по нескольким атрибутам, каждый из которых допускает экспертную оценку. Задача эксперта выбрать один из индикаторов соответствия данных критериям качества. В качестве типовых критериев при сертификации принимают такие факторы как обоснованность (*validity*), полнота, согласованность, однородность и ряд других, с конкретизацией требований, предъявляемых в соответствии с данным критерием.

Сертификация, предложенная в работе [17] для оценивания данных по физическим свойствам вещества, основана на использовании трех атрибутов или критериев качества: (1) полнота информации о состоянии и подготовке образца; (2) полнота описания метода и приборов, используемых при измерении (или кодов и техники обработки данных); (3) согласованность численных данных с известными закономерностями или ранее полученными надежными результатами. В сочетании с оценкой неопределенности, подобная сертификация проверяет выполнение трех важнейших критериев качества: неопределенности, полноты и согласованности. В ходе сертификации эксперт приписывает каждому из атрибутов качества соответствующий индикатор (высокий, средний или низкий), определяющий соответствие данных предъявляемым требованиям.

Наличие трех оценок при сертификации вместе с оценкой неопределенности означает, что качество набора данных аттестовано с достаточной полнотой и нуждается в коррекции лишь при получении новых данных с более высокой достоверностью.

Назначение метаданных. Для любых цифровых данных, вне зависимости от происхождения и структуры, важнейшим из условий их долгосрочного хранения является наличие метаданных, т. е. структурированной информации, призванной описывать состав, происхождение, структуру и другие признаки набора данных. Без них набор данных сводится просто к последовательности битов без возможной интерпретации. Процесс курирования включает как назначение метаданных с момента создания цифрового архива, так и их постоянное использование на важнейших стадиях хранения, когда проводится смена форматов, среды хранения, а также дополнительное резервирование. Специальный вид метаданных, так называемые метаданные о хранении (*preservation metadata*), отражают информацию об условиях хранения, включая сведения о том, как данные были получены, когда и кем введены в архив, каким процедурам обработки и ревизии подвергались за период хранения [18].

Курирование позволяет использовать метаданные для включения контекста, т. е. для связывания набора данных с тематически близкими ресурсами в том же архиве или внешней среде: репозитариях, БД, *web*-порталах и т.п. Метаданные включают для этой цели специальные элементы, предназначенные для устойчивых идентификаторов (*DOI*, *URL* и т.п.), определяющих локализацию адекватного ресурса. Тем самым, регулярно проводимое курирование позволяет обновлять данные без физического изъятия, нежелательного из-за их исторической ценности. Другая возможность усиления ценности данных по мере хранения основана на метаданных, отражающих их качество. Поскольку по мере появления новых данных, полученных в результате исследований, может измениться представление о точности и согласованности прежних, куратор занижает для них оценки качества, одновременно включая ссылки на идентификатор с новыми результатами.

По метаданным, которые документируют и аннотируют наборы научных данных, предоставляя вспомогательную информацию, необходимую для их поиска, интерпретации, оценки и использования, имеется ряд публикаций, например, [19, 20]. Стандарты метаданных для разных дисциплин и типов документа собраны в *Metadata Standard Directory* (MSD, http://rd-alliance.github.io/metadata-directory) и Центре цифрового курирования. Там же имеются стандарты, независимые от выбора дисциплины, например, *Dublin Core metadata set* (http://dublincore.org/) для формального описания цифровых ресурсов или *DataCite Metadata Store* (http://schema.datacite.org) для их идентификации и цитирования. В работах [5, 21, 22] детально описаны метаданные для теплофизических свойств (*ThermoML*), характеристик обычных материалов (*MatML*) и объектов нанотехнологий.

Вне зависимости от предметной области, научные метаданные должны удовлетворять ряду требований, гарантирующих достаточную полноту и достоверность в представлении набора данных посредством: однозначной идентификации объекта исследования; предоставления сведений об источнике данных (метод исследования, оборудование, программный код и т.п.); информации о неопределенности и качестве данных; связи с контролируемыми словарями или онтологиями; возможности гибкой подстройки к особенностям объекта и его характеристикам.

Опубликовано специальное руководство по научным метаданным и их использованию в процессе курирования [20]. В качестве наиболее распространенного там выделен тип метаданных в виде

[4] В БД величина, обозначенная как NA (или Not Applicable, Null), означает полное отсутствие величин в некотором поле, т. е. что величина в поле остается неизвестной - см. www.techopedia.com/definition/5539/null

пары «имя-значение», в которой элементу информации, идентифицируемому по имени, присваивается определенное значение. Примером такой пары может служить любой из приведенных выше атрибутов качества данных, например «согласованность», которому присваивается одно из трех значений индикатора (высокий, средний или низкий). Другой пример связан с детализацией понятия о неопределенности данных, для чего в формате теплофизических данных *ThermoML* [19, 21] предусмотрено три вида оценок: стандартные σ, расширенные[5] и комбинированные, объединяющие вклады зависимой и независимой переменных. Тем самым, в паре «имя-значение» понятие «неопределенность» можно рассматривать как имя, которому присваивается одно из трех значений: стандартные, расширенные, комбинированные.

В приведенных примерах значение выбирается из предварительно составленного списка. В более общем случае для этой цели используется контролируемый словарь с обширным множеством понятий, каждое из которых определяется идентификатором (ID). Так, для точной идентификации вещества с учетом структуры и изомерии используется номер CAS (*Chemical Abstract Service*) в виде трёх групп арабских чисел, разделённых дефисами (для H_2O, например, это номер – 7732-18-5).

Заметим, что в паре «имя-значение» можно, наряду с терминами контролируемого словаря, использовать произвольный текст и числа (целые или с плавающей запятой). При этом, если значение в этой паре не является числовым, элементом перечислимого набора или устойчивым идентификатором (типа номера CAS или URL), то оно полезно только для восприятия человеком-оператором.

Характер метаданных, привлекаемых для аннотирования научной информации, иллюстрирует типовая запись в репозитарии Института стандартов и технологий США (NIST Materials Data Repository, https://materialsdata.nist.gov/), где собраны наборы экспериментальных и расчетных данных, полученных при подготовке статей по свойствам материалов (рис. 1 и 2).

Метаданные на рис. 1 включают уникальный идентификатор URI (сетевой адрес), перечень файлов с указанием для каждого имени, формата и размера. Для того же набора данных дается развернутая система метаданных в стандарте *Dublin Core* (http://dublincore.org/), который предназначен для идентификации произвольного ресурса в сети (рис. 2). Пятнадцать базовых элементов предоставляют необходимые сведения об авторстве, дате выпуска, локализации, правовых аспектах и т.п. Стандарт не имеет какой-либо тематической ориентации, но он достаточно прост и эффективен при кратком описании информационного ресурса, благодаря чему приобрел статус международного (ISO Standard 15836). Все метаданные из набора построены по принципу «имя-значение», причем большая часть значений имеет вид произвольного текста, за исключением дат и адресов URI для набора данных и лицензии, и только один элемент *dc.subject* дает сведения о содержании набора данных посредством ключевых слов (см. рис. 2), но без всякой ссылки на контролируемые словари. Заметим, что здесь слабое отражение в метаданных специфики предметной области оправдано, поскольку сами данные дополняют публикацию, которая содержит необходимые сведения по свойствам изучаемого материала, методам исследования и представлению результатов.

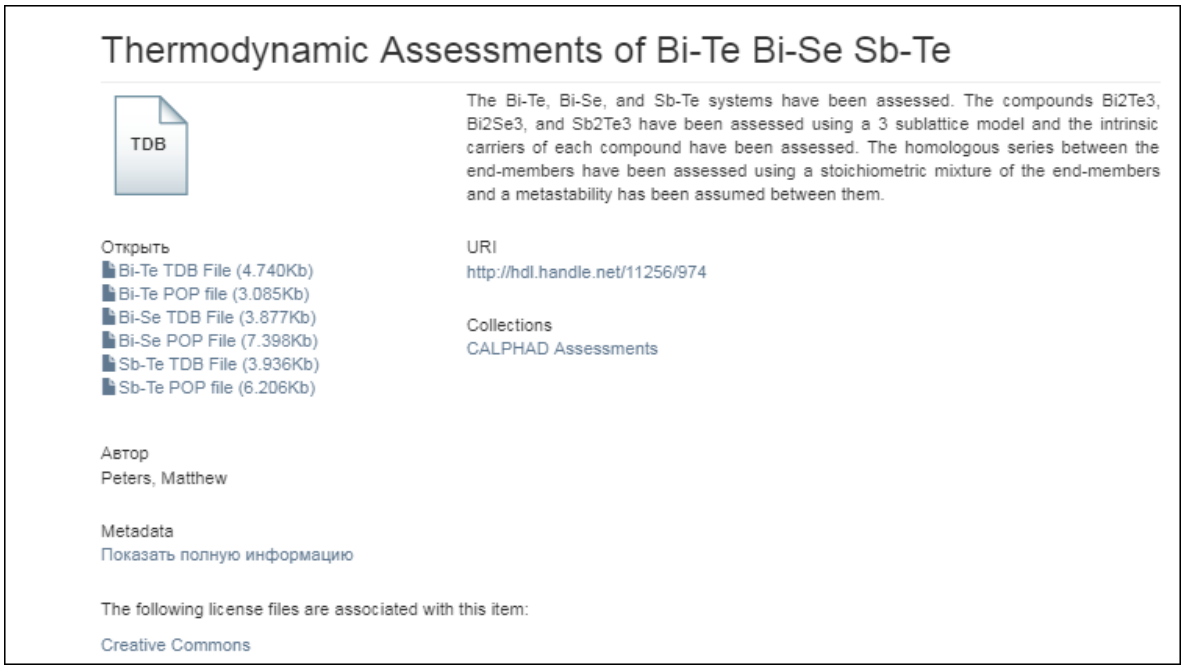
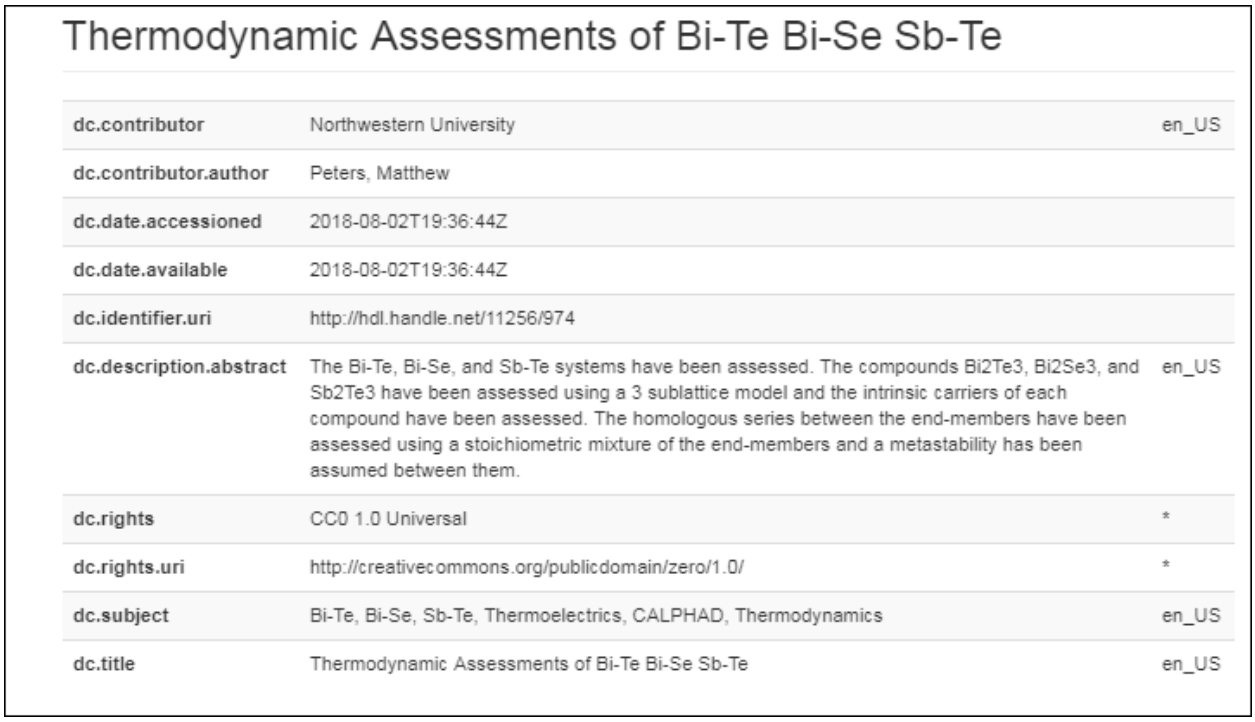


Рис. 1. Термодинамические данные для бинарных сплавов Bi-Te, Bi-Se, Sb-Te.
Скан записи на портале NIST (https://materialsdata.nist.gov/).

[5] Расширенная неопределенность $\sigma_L = k\sigma$ определяется по уровню значимости, с которым однозначно связан фактор k. Как правило, в качестве уровня значимости принимается 95%, чему соответствует $k \approx 2$.



Thermodynamic Assessments of Bi-Te Bi-Se Sb-Te

dc.contributor	Northwestern University	en_US
dc.contributor.author	Peters, Matthew	
dc.date.accessioned	2018-08-02T19:36:44Z	
dc.date.available	2018-08-02T19:36:44Z	
dc.identifier.uri	http://hdl.handle.net/11256/974	
dc.description.abstract	The Bi-Te, Bi-Se, and Sb-Te systems have been assessed. The compounds Bi2Te3, Bi2Se3, and Sb2Te3 have been assessed using a 3 sublattice model and the intrinsic carriers of each compound have been assessed. The homologous series between the end-members have been assessed using a stoichiometric mixture of the end-members and a metastability has been assumed between them.	en_US
dc.rights	CC0 1.0 Universal	*
dc.rights.uri	http://creativecommons.org/publicdomain/zero/1.0/	*
dc.subject	Bi-Te, Bi-Se, Sb-Te, Thermoelectrics, CALPHAD, Thermodynamics	en_US
dc.title	Thermodynamic Assessments of Bi-Te Bi-Se Sb-Te	en_US

Рис. 2. Набор метаданных в формате **Dublin Core.**
Скан записи на портале NIST (https://materialsdata.nist.gov/).

Таблица 2

Метаданные, предназначенные для сопровождения результатов использования метода CALPHAD

Тип материала	Объемный состав
	Чистота материала
	Подготовка образца
	Информация о микроструктуре
	Данные для монокристалла
	Данные для поликристалла (размер граней, плотность дислокаций)
	Данные для некристаллической фазы
Детали работы с данными	Поправки к базисному состоянию
	Методы определения *интердиффузионного коэффициента*
Формат данных	
Ссылки	DOI
	file

ПРИМЕЧАНИЕ. Понятие об интердиффузионном коэффициенте связано с так называемой химической диффузией в твердых растворах, когда движение одного компонента вызывает противоток другого, или вакансий [25].

В ряде случаев метаданные обеспечивают значительно большую детализацию содержания. Так, в проекте репозитария материаловедческих данных [23], создаваемого для будущей генерации вычислительной технологии CALPHAD [24], на метаданные возлагается детализация материала и метода исследований [25] (табл. 2).

При хранении данных компьютерной обработки эксперимента на метаданные возлагается еще одна задача [20]. В метаданных тогда приходится отражать тот факт, что в реальной практике используются программы, созданные без привлечения профессиональных программистов. Каждая из таких программ похожа на ранее неизвестный образец прибора, к тому же не прошедшего стадии рецензирования. Если в прошлом программное обеспечение просто ускоряло обработку данных, то сейчас из второстепенного аксессуара оно превратилось в инструмент, по значимости соизмеримый с основным научным оборудованием. В этом случае роль метаданных сводится к представлению плохо формализуемой информации с указанием местонахождения и состава нестандартного программного обеспечения, документации и правил использования, нестандартных форматов данных.

Проблемы, вызванные «любительским» уровнем научного софта, давно привлекали к себе внимание специалистов по программной инженерии [26–28]. Например, один из блогов компьютерного писателя Shannon Love так и называется «Ученые не программисты» [28]. По существу, указанная проблема не может быть решена в рамках курирования данных, поскольку затрагивает всю методологию современных исследований. Роль метаданных здесь относительно скромна – они даже не описывают нестандартный софт, созданный в процессе исследования, а скорее регламентируют требования, которые должны гарантировать полноту описания и возможности повторного использования стареющего программного обеспечения. В дальнейшем переход к промышленному программному обеспечению должен сочетаться с хранением «любительских» копий для проверки воспроизводимости результатов.

СТАНДАРТЫ КУРИРОВАНИЯ

Без привлечения стандартов, регламентирующих требуемые действия куратора и их последовательность, курирование обширных цифровых архивов немыслимо. В принципе, стандартизация позволяет выдерживать единый порядок функционирования архивов и возможность согласованной работы при их интеграции. Однако в справочнике [4] отмечено, что есть ряд трудноразрешимых проблем с широким применением стандартов:

- быстрая эволюция программно-аппаратного обеспечения, еще не отраженная в действующих стандартах;
- непатентованные форматы файлов, жестко связанные с определенным видом софта и методами моделирования или обработки данных;
- различия в стиле представления и полноте описания данных.

Иллюстрацией к последнему пункту служат различия в метаданных репозитария NIST и репозитария CALPHAD при том, что оба хранилища предназначены для однотипных данных (см. рис. 1, 2 и табл. 2). Авторы [4] специально подчеркивают, что стандарты могут рассматриваться лишь как элемент в стратегии курирования, которая не может полностью на них основываться. Практика показывает, что оптимальной является замена жестких стандартов на более гибкие подходы, одним из которых является, так называемая «наилучшая практика» (*best practice*), которая ориентирует на предпочтительное использование лишь некоторых элементов курирования в условиях изменяющейся техники и профессиональных требований к данным. К таким элементам справочник [4] относит всего три: использование открытых, непатентованных форматов данных; согласование метаданных с новыми стандартами; присвоение устойчивых имен (*persistent name*) отдельным наборам данных [29].

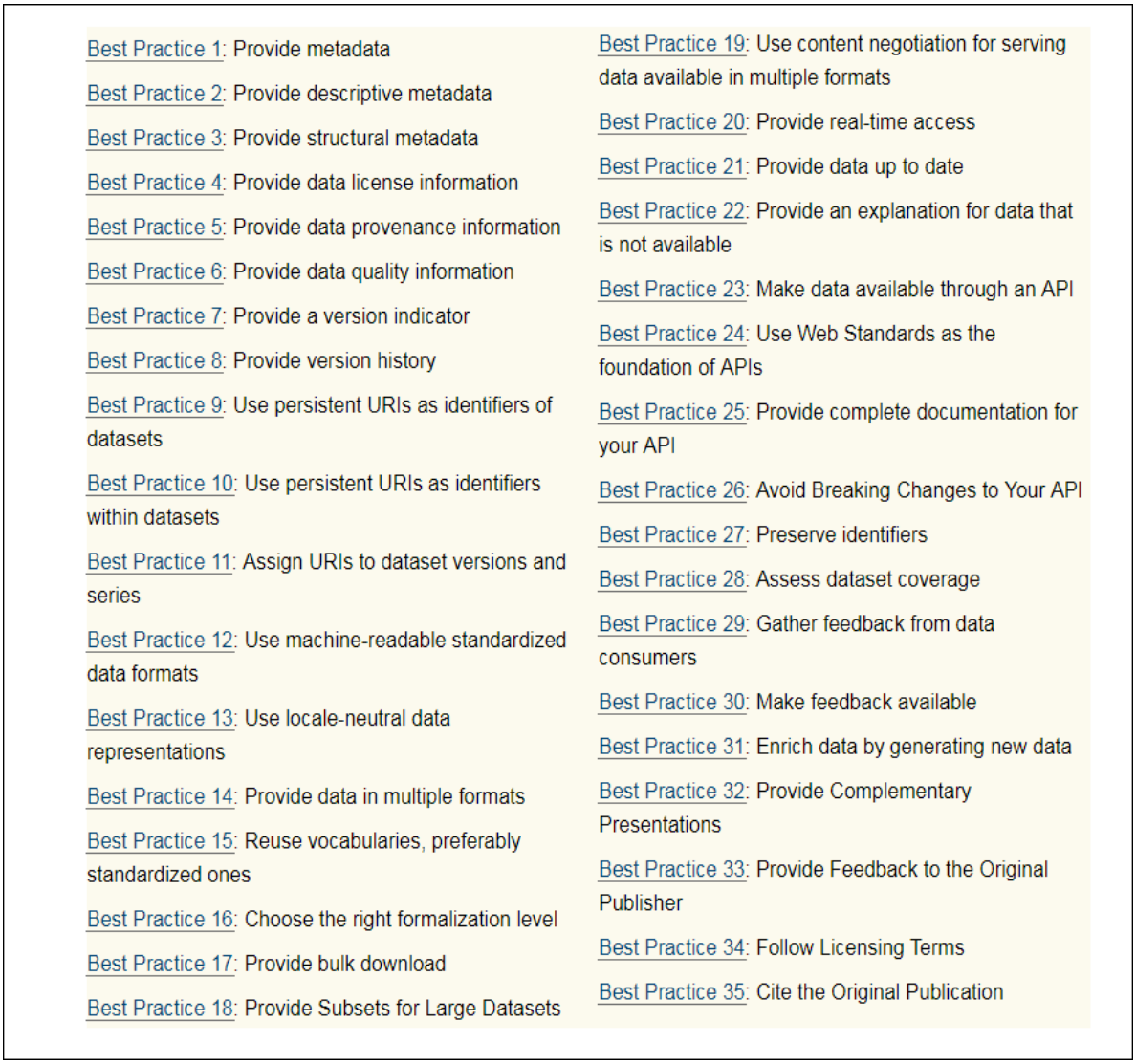
Подготовленный консорциумом W3C[6] нормативный документ [30] предлагает набор из 35 «наилучших практик», обеспечивающих представление данных в сети (рис. 3). Документ охватывает большинство мер, предусмотренных в процессе курирования: назначение детализированных метаданных (п.п. 1-6), обязательность информации о происхождении и качестве данных (п.п. 5 и 6), обогащение данных (п. 31). Поскольку, однако, документ [30] напрямую не предназначен для курирования, а ограничен публикацией и использованием данных, то в нем не предусмотрено стандартных мер по сохранности, очистке данных или миграции форматов.

По-видимому, наибольший набор из 47 «наилучших практик», охватывающих все аспекты курирования данных, предложен в работе [3] (см. раздел «**Основные составляющие процесса курирования»)**. В табл. 1 приведены основные меры по долгосрочному хранению цифровых данных. Помимо них, в этом наборе предусмотрены многообразные меры по очистке данных, правовому регулированию (ограничения доступа, лицензионные условия, указатели персональной или конфиденциальной информации), контролю и фиксации изменений с момента ввода данных и т.д. Наиболее важные для архива меры описания научных данных приведены в табл. 3. Среди них редактирование метаданных (например, рубрики «Индексирование» и «Контекстуализация»), регулярный контроль их полноты, использование в сетевых поисковых системах. Особую роль играет, так называемая, *контекстуализация*, т. е. включение собственных данных в более общий контекст за счет связи с тематически родственными ресурсами. Это позволяет реализовать важнейшую задачу процесса курирования: обогащение исходных данных за счет их связывания с новыми данными, появившимися в репозитарии или Web.

Стандарты метаданных. Курирование данных, предусмотренное «наилучшими практиками» [3, 30], оставляет достаточный уровень свободы в реализации конкретных целей и, прежде всего, в назначении и использовании метаданных. В табл. 3, например, при определении понятий «Индексирование» и «Интероперабельность» содержатся лишь намеки на желательность согласования со стандартами. Поэтому общие рекомендации по курированию данных приходится дополнять независимыми разработками стандартов метаданных [4], среди которых можно выделить, в основном, две группы. Стандарты первой группы (*preservation metadata*) документируют условия хранения, т. е. отчетность о действиях с цифровым материалом для его поддержки в течение длительного времени [4, 18]. Вторая группа стандартов (*disciplinary metadata*) имеет непосредственное отношение к науке, документируя базовые понятия каждой из дисциплин в соответствии с принятыми в ней словарями терминов.

Стандарты первой группы созданы, в основном, для управления обширными цифровыми библиотеками и репозитариями, вне зависимости от тематики и вида документов, начиная с обычных текстов и вплоть до аудио- или видео продукции. По существу, они детализируют отчетность по мерам долговременного хранения, очерченным наилучшими практиками [3, 4, 30].

[6] W3C – организация, разрабатывающая единые принципы и стандарты для глобальной сети. Внедрение W3C рекомендаций обеспечивает совместимость программно-аппаратного обеспечения, действующего в сети.



Best Practice 1: Provide metadata
Best Practice 2: Provide descriptive metadata
Best Practice 3: Provide structural metadata
Best Practice 4: Provide data license information
Best Practice 5: Provide data provenance information
Best Practice 6: Provide data quality information
Best Practice 7: Provide a version indicator
Best Practice 8: Provide version history
Best Practice 9: Use persistent URIs as identifiers of datasets
Best Practice 10: Use persistent URIs as identifiers within datasets
Best Practice 11: Assign URIs to dataset versions and series
Best Practice 12: Use machine-readable standardized data formats
Best Practice 13: Use locale-neutral data representations
Best Practice 14: Provide data in multiple formats
Best Practice 15: Reuse vocabularies, preferably standardized ones
Best Practice 16: Choose the right formalization level
Best Practice 17: Provide bulk download
Best Practice 18: Provide Subsets for Large Datasets
Best Practice 19: Use content negotiation for serving data available in multiple formats
Best Practice 20: Provide real-time access
Best Practice 21: Provide data up to date
Best Practice 22: Provide an explanation for data that is not available
Best Practice 23: Make data available through an API
Best Practice 24: Use Web Standards as the foundation of APIs
Best Practice 25: Provide complete documentation for your API
Best Practice 26: Avoid Breaking Changes to Your API
Best Practice 27: Preserve identifiers
Best Practice 28: Assess dataset coverage
Best Practice 29: Gather feedback from data consumers
Best Practice 30: Make feedback available
Best Practice 31: Enrich data by generating new data
Best Practice 32: Provide Complementary Presentations
Best Practice 33: Provide Feedback to the Original Publisher
Best Practice 34: Follow Licensing Terms
Best Practice 35: Cite the Original Publication

Рис. 3. Скан из документа [30] с перечнем из 35 «наилучших практик», рекомендованных консорциумом W3C.

Таблица 3

Меры и средства эффективного описания наборов данных [3]

Метаданные	Информация о наборе данных, структурированная (часто в машиночитаемом формате) для целей поиска. Элементы метаданных могут включать в себя основную информацию (например, название, автора, дату создания) и/или конкретные элементы, свойственные наборам данных (например, пространственный или временной интервал).
Индексирование	Дополнение авторских метаданных описательными и административными метаданными, совместимыми со стандартами, принятыми в репозитарии.
Интероперабельность	Форматирование данных с использованием дисциплинарных стандартов для лучшей интеграции с другими наборами данных и/или системами.
Распространение метаданных	Активное распространение метаданных в службы поиска, например, в базы данных статей, каталоги, *web*-индексы для федеративного поиска и обнаружения.
Документация	Хранение любой необходимой информации для использования и понимания данных. Документация может быть структурирована (книга кодов) или неструктурированна, например, в виде текстового файла “Readme”.
Контекстуализация	Использование метаданных, чтобы связать набор данных с тематически родственными материалами – публикациями, диссертациями, проектами, которые предоставляют дополнительный контекст для понимания, того, как и почему были получены данные.
Гарантия качества	Контроль полноты и достоверности документации и метаданных.

Наибольшей универсальностью среди них обладают PREMIS (Preservation Metadata: Implementation Strategies), разработанный OCLC[7], и METS (*Metadata Encoding and Transmission Standard*, www.loc.gov/standards/). Если в стандарте PREMIS исключены те метаданные, которые непосредственно не связаны с процессом хранения, например детали информации об аппаратном окружении и среде хранения, то стандарт METS позволяет *упаковывать* цифровые материалы вместе с архивной информацией, включающей административные, описательные и структурные метаданные[8]. Подробное описание этих и других стандартов, определяющих режим хранения цифровых документов, можно найти в справочнике [4] и обзорах [18, 31].

Применительно к научному контенту (данным эксперимента, моделирования), наряду с контролем сохранности, требуется глубокая систематизация материала, обеспечивающая возможности поиска и повторного использования. С этой задачей призваны справиться стандарты второй группы (дисциплинарные метаданные), каталог которых представлен в директории MSD (*Metadata Standards Directory*, http://rd-alliance.github.io/metadata-directory). История и принципы его разработки анализируются в статье [32]. За основу был принят аналогичный каталог Центра DCC (www.dcc.ac.uk/resources/metadata-standards), подвергнутый обновлению, расширению и переносу на новую платформу. Собранные в директории MSD стандарты должны были определять подробные описательные метаданные, предписывающие, какую информацию о данных собирать, причем для преимущественного документирования табличных данных. В то же время было принято решение по возможности исключать стандарты:

- предписывающие как структурировать или передавать данные и метаданные;
- сосредоточенные на администрировании, сохранении или более широком круге действий;
- описывающие публикации, учебные объекты, аудиовизуальные файлы или повествовательный текст (например, стенограммы интервью).

Каталог MSD охватывает 5 предметных областей (*Arts and Humanities, Engineering, Life Sciences, Physical Sciences & Mathematics, Social and Behavioral Sciences*), разбитых на отдельные дисциплины. Для каждой из дисциплин в каталоге указаны 4 коллекции ресурсов: стандарты метаданных, расширения, программные средства, варианты использования. Под термином «расширения» понимаются профили приложений, которые адаптируют стандарт к определенным типам репозитариев или данных; к «средствам» относятся сервисы и программы для работы с конкретным стандартом, например, редактор метаданных. Наконец, к «вариантам использования» каталог относит известные примеры репозитариев или библиотек, где активно используется данный стандарт. Наряду с пятью предметными областями, каталог содержит и мультидисциплинарные исследования (*General Research Data*), включающие указанные типы ресурсов.

Например, запись на нижнем уровне каталога для определенного стандарта *Crystallographic Information Framework* (CIF) включает краткое описание стандарта, ссылки на связанные со стандартом ресурсы (сайт, спецификацию, родственные словари), перечни областей применения, а также ссылки на другие записи каталога: для расширений, средств и вариантов использования.

Несмотря на обширный перечень в MSD стандартов, охватывающих даже узкие сферы (например, «Гляциология» из области *Physical Sciences & Mathematics* или «Рыболовство» из области *Life Sciences*), трудно признать, что на сегодняшний день они адекватно отражают потребности в систематизации, во всяком случае, в области физико-математических или химических дисциплин. Для этого достаточно взглянуть на стандарты, отнесенные в каталоге к таким базовым дисциплинам как, например, «Физика» или «Материаловедение». Например, к дисциплине «Физика» в каталоге приписан один тематический стандарт PDBx/mmCIF (*Protein Data Bank Exchange Dictionary and the Macromolecular Crystallographic Information Framework*), а к материаловедению стандарт CIF. Поэтому создатели многочисленных баз данных и репозитариев по свойствам веществ и материалов вынуждены пользоваться собственными наборами метаданных или словарей. Например, в репозитарии NIST (см. рис. 1 и 2) вообще отказались от тематических метаданных в пользу универсального стандарта *Dublin Core* (http://dublincore.org/), определяющего произвольный ресурс без учета его специфики и предметной области. Для тематически близкого репозитария CALPHAD приняты собственные метаданные вне всякой связи с существующими стандартами (см. табл. 2).

Вообще, стандартизация метаданных для сколько-нибудь широкой тематики представляется проблематичной. Особенно заметной оказалась эта проблема в науке о материалах, многообразие которых, наряду с обширной номенклатурой характеристик и технологий, казалось бы, исключает возможность единой схемы данных. Так, Д.Е. Бойс и др. [33] замечают, что метаданные всегда зависят от деталей проекта, который может фокусироваться на различных характеристиках материала. По этому же поводу авторы другой работы [34] отмечают невозможность поддерживать на основе единого стандарта сеть из БД для различных материалов.

Поскольку, однако, отказ от стандартов влечет известные проблемы с интеграцией, идут поиски нестандартных подходов к объединению метаданных тематически разнородных ресурсов [5]. Один из таких подходов базируется на стандартах, которые не имеют четко выраженной тематической направленности. Примером служит упомянутый репозитарий NIST, использующий стандарт *Dublin Core*, ограниченный при тематическом поиске набором

[7] *Online Computer Library Center* (www.oclc.org) – международная исследовательская организация и сервис, обеспечивающие расширение доступа к мировой информации.

[8] На сайте Президентской библиотеки им. Б.Н. Ельцина представлены переводы на русский язык основных документов по стандарту METS. – URL: www.prlib.ru/mets.

ключевых слов. Однако каталог MSD включает стандарты с более эффективными возможностями. Среди них CSMD (*Core Scientific Metadata Model*, http://icatproject-contrib.github.io/CSMD/), легко адаптируемый к разным дисциплинам. Стандарт разрабатывался на протяжении последних десятилетий преимущественно для проектов, выполняемых на масштабных установках типа нейтронных источников, телескопов [35]. Предусматривалась и возможность более широкого его использования, особенно в таких дисциплинах как химия, материаловедение, науки о Земле и др.

Модель CSMD формирует иерархическую структуру научных исследований в виде экспериментов, наблюдений, измерений и моделирования. Их результаты образуют наборы данных (сырых, т. е. необработанных), проанализированных и прошедших обработку) и итоговых результатов. Наборы данных группируются в соответствии с параметрами, определяющими условия исследования. При этом, модель не содержит каких-либо ограничений в терминологии, так что может использоваться для различных приложений и позволяет интегрировать распределенные гетерогенные метаданные в однородную платформу.

В качестве базовых стандарт включает следующие элементы: *Investigation (исследование), Investigator (исследователь), Topic and Keyword (тематика и ключевое слово), Sample (образец), Publication (публикация), Dataset (набор данных), Datafile (файл данных), Parameter (параметр), Authorisation (авторизация)*. Каждый из них может иметь несколько полей для детализации. Например, элемент *Investigation* включает название, аннотацию, даты (начала, окончания исследования), регистрационный номер, перечень оборудования и инструментов. Другой элемент *Sample* содержит поля для идентификации объекта исследования, например химическую формулу, данные о размере, форме, структуре. Существенно также, что элемент *Topic and Keyword* может иметь ссылки на контролируемые словари, что однозначно определяет понятия, по которым проводится поиск. Так, компендиум химической терминологии **Goldbook** закрепляет за каждым из терминов сетевой адрес (URL), например https://goldbook.iupac.org/html/S/S05915.html за понятием "*standard equilibrium constant*". Привлекая совокупность словарей и/или онтологий, можно достаточно полно отразить тематические рамки, включая объект, его характеристики, методы исследования и т.п. Ссылки на контролируемые словари в сочетании с детализацией элемента *Sample* открывают возможности поиска по множеству критериев, во всяком случае, в большинстве естественнонаучных дисциплин.

Наряду с детализированным описанием набора данных, стандарт **CSMD** предоставляет богатые возможности для хранения массивов данных, полученных в итоге исследования. Наборы данных рассматриваются как иерархии, которые могут содержать поднаборы, в свою очередь разбиваемые на отдельные файлы логических данных, которые могут храниться как в хранилище, так и в БД. На каждом уровне детализации метаданные дают информацию о физическом расположении. Поскольку в модели различается логическое хранение и расположение данных, средствами метаданных обеспечивается связь идентификаторов определения данных с фактическими URL-адресами файлов. Тем самым, участники проекта, как и посторонние эксперты получают принципиальную возможность доступа к результатам любого этапа исследования с различением сырых, промежуточных и итоговых данных. Сверх этого элемент *Authorisation* в наборе CSMD может регламентировать возможности доступа к каждому из наборов данных.

ЗАКЛЮЧЕНИЕ

В настоящей статье мы показали, что рутинные действия по сохранности данных, на что исходно нацелен процесс курирования, применительно к масштабным цифровым архивам, приобретают характер серьезной научно-методической работы. Прежде всего, цифровые ресурсы жестко связаны с текущим программно-аппаратным обеспечением, и особенно с принятыми форматами записей. Поэтому в процессе курирования необходимо отслеживать не только состояние архива, но и доступность новых средств для преодоления технологического старения (*technological obsolescence* [4]). Многочисленные проблемы, сопровождающие долгосрочное хранение данных, потребовали разработки специальных руководств [3, 4, 18] по выбору мер и средств, предупреждающих технологическое устаревание среды хранения и форматов. Среди наиболее важных мер – принятие определенных стандартов, регламентирующих долгосрочное хранение и документирование данных, что необходимо для согласованной деятельности различных хранилищ (репозитариев и БД).

Если технические меры по долгосрочному хранению связаны, в основном, с объемом и форматом архивного материала, то курирование научных данных напрямую связано с исследованиями, в ходе которых они созданы. Проявляется эта связь в метаданных, отражающих свойства объекта и методы его исследования, присвоение атрибутов качества, связывание данных с родственными ресурсами: публикациями, *web*-страницами. В стратегии долгосрочного хранения появляется возможность сократить объем данных, если по требованию пользователя доступна процедура их реконструкции из исходных. Как правило, речь идет о воспроизведении данных компьютерного моделирования при наличии информации о программном обеспечении и исходных параметрах.

Последовательное использование куратором различных средств хранения, описания и оценивания научных данных гарантирует не только их сохранность, но и перманентное обогащение содержания, например, при ссылках на более достоверные данные, и сопутствующую переоценку качества. Наконец, новые методы курирования открывают путь к решению давней проблемы: как сочетать детализированное описание конкретной области знания с устранением искусственных барьеров для обнаружения и повторного использования данных в разных дисциплинах.

СПИСОК ЛИТЕРАТУРЫ

1. Ball A. Review of the State of the Art of the Digital Curation of Research Data. (Version 1.1). ERIM Project Document erim1rep091103ab11. – Bath, UK: University of Bath, 2010.
2. Palmer C., Weber N., Muñoz T., Renar A. Foundations of Data Curation: The Pedagogy and Practice of "Purposeful Work" with Research Data // Archives Journal. – 2013. – Vol. 3. – URL: http://hdl.handle.net/2142/78099
3. Johnson L.R., et al. How Important Are Data Curation Activities to Researchers? Gaps and Opportunities for Academic Libraries// Journal of Librarianship and Scholarly Communication. –2018 – 6(General Issue). – eP2198. – URL: https://doi.org/10.7710/2162-3309.2198
4. Preservation Management of Digital Materials: The Handbook. Digital Preservation Coalition. 2008. – URL: www.dpconline.org/graphics/handbook/
5. Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А. Интенсивное использование цифровых данных в современном естествознании // Научно-техническая информация. Сер. 2. – 2017. – №9. – С. 9-22; Erkimbaev A.O., Zitserman V.Yu., Kobzev G.A. The Intensive Use of Digital Data in Modern Natural Science // Automatic Documentation and Mathematical Linguistics. – 2017. – Vol. 51, № 5. – P. 201-213.
6. The Fourth Paradigm. Data-Intensive Scientific Discovery / ed. by T. Hey, St. Tansley, and Kr. Tolle. Microsoft Corporation – 2009.
7. Kingsbury D., Snoddy J., Robbins R. Report of the Invitational DOE Workshop on genome informatics, 26-27 April 1993, Baltimore, Maryland. Genome Informatics I: Community Databases // Journal of Comparative Biology. – 1994. – Vol. 1. – P. 173–190.
8. Zorich D.M. Data management: managing electronic information: data curation in Museums // Museum Management and Curatorship. – 1995. – Vol. 14(4). – P. 430–432.
9. Beagrie N., Pothen P. Digital Curation: digital archives, libraries and e-science seminar// Ariadne. – 2001. – Issue 30. – URL: www.ariadne.ac.uk/issue30/digital-curation/
10. Lord P., Macdonald A. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision // The JISC Committee for the Support of Research (JCSR). 2 Wayside Court, Arlington Road, Twickenham, TW1 2BQ. The Digital Archiving Consultancy Limited. 2003.
11. Abbott D. What is Digital Curation? DCC Briefing Papers. Introduction to Curation. Edinburgh: Digital Curation Centre, 2008. – URL: www.dcc.ac.uk/resources/briefing-papers/introduction-curation
12. Gray J., Szalay A.S., Thakar A.R. et al. Online Scientific Data Curation, Publication, and Archiving // Technical Report MSR-TR-2002-74. – Redmond, WA 98052: Microsoft Research, 2002.
13. Osborne J.W. Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data // Newborn and Infant Nursing Reviews. – 2010. – Vol. 10, Iss. 1. – P. 37-43.
14. Rahm E., Do H.H. Data cleaning: Problems and current approaches // IEEE Data Eng. Bull. – 2000. – Vol. 23, № 4. – P. 3 –13.
15. Elgabry O. The Ultimate Guide to Data Cleaning. – URL: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4
16. Enders C.K. Applied Missing Data Analysis. – New York: Guilford Press, 2010.
17. Елецкий А.В., Еркимбаев А.О., Зицерман В.Ю. и др. Теплофизические свойства наноразмерных объектов: систематизация и оценка достоверности данных // Теплофизика высоких температур. – 2012. – Т. 50, №4. – С. 524 – 532.
18. Caplan P. Preservation Metadata / eds. S. Ross, M. Day. DCC Digital Curation Manual, 2006. – URL: www.dcc.ac.uk/resource/curation-manual/chapters/preservation-metadata
19. Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А. Роль метаданных в создании и использовании информационных ресурсов о свойствах веществ и материалов // Научно-техническая информация. Сер. 1. – 2008. – № 11. – С. 13-19; Yerkimbaev A.O., Zitserman V.Yu., Kobzev G.A. The Role of Metadata in the Creation and Application of Information Resources on the Properties of Substances and Materials // Scientific and Technical Information Processing. – 2008. – Vol. 35, № 6. – P. 247-255.
20. Davenhall C. Scientific Metadata / eds. J. Davidson, S. Ross, M. Day. DCC Digital Curation Manual, 2011. – URL: www.dcc.ac.uk/resources/curation-reference-manual/scientific-metadata
21. Frenkel M. Global communications and expert systems in thermodynamics: Connecting property measurement and chemical process design// Pure Applied Chem. – 2005. – Vol. 77, № 8. – P. 1349 – 1367.
22. Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Трахтенгерц М.С. Универсальная система метаданных для характеристики наноматериалов // Научно-техническая информация. Сер. 1. – 2015. – №10. – С. 8-20; Erkimbaev A.O., Zitserman V.Yu., Kobzev G.A., Trakhtenhers M.S. A Universal Metadata System for the Characterization of Nanomaterials // Scientific and Technical Information Processing. – 2015. – Vol. 42, № 4. – P. 211-222.
23. Campbell C.E., Kattner U.R., Liu Z.-K. File and data repositories for Next Generation CALPHAD // Scripta Materialia. – 2014. – Vol. 70. – P. 7–11.
24. Kaufman L., Bernstein H. Computer Calculation of Phase Diagrams. – London: Academic Press, 1970.
25. Definitions of terms for diffusion in the solid state (IUPAC Recommendations 1999) // Pure Appl. Chem. –1999. – Vol. 71, №7. – P. 1307–1325.

26. Goble C., De Roure D. Curating Scientific Web Services and Workflow // EDUCAUSE Review. – 2008. – Vol. 43, № 5. – P. 10-11.
27. No One Peer-Reviews Scientific Software. Posted by Shannon Love on November 28th, 2009. – URL: https://chicagoboyz.net/archives/10436.html
28. Scientist Are not Software Engineers. Posted by Shannon Love on November 28th, 2009. – URL: https://chicagoboyz.net/archives/10399.html
29. Klump J. et al. Editorial: 20 Years of Persistent Identifiers – Applications and Future Directions // Data Science Journal. – 2017. – Vol. 16. – Art. # 52. – P. 1–7
30. Data on the Web Best Practices. W3C Recommendation 31 January 2017. – URL: www.w3.org/ TR/dwbp/
31. Day M. Metadata / eds. S. Ross, M. Day. DCC Digital Curation Manual, 2005. – URL: www.dcc.ac.uk/ resource/curation-manual/chapters/ metadata/
32. Ball A. et al. Building a Disciplinary Metadata Standards Directory // Int. Journ. of Digital Curation. – 2014. – Vol. 9, Iss. 1. – P. 142–151.
33. Boyce D.E., Dawson P.R., Miller M.P. The Design of a Software Environment for Organizing, Sharing, and Archiving Materials Data // Metallurgical and Materials Transactions A. – 2009. – Vol. 40A. – P. 2301– 2318.
34. Surya R. Kalidindi and Marc De Graef. Materials Data Science: Current Status and Future Outlook // Annu. Rev. Mater. Res. – 2015. – Vol. 45. – P. 171–193
35. Matthews B. et al. Using a Core Scientific Metadata Model in Large-Scale Facilities // Int. Journ. of Digital Curation. – 2010. – Vol. 5, Iss.1. – P. 106 – 118.

Материал поступил в редакцию 24.06.19

Сведения об авторах

ЕРКИМБАЕВ Адильбек Омирбекович – кандидат технических наук, зам. зав. лабораторией теплофизических баз данных Объединенного института высоких температур Российской академии наук (ОИВТ РАН), Москва
e-mail: adilbek@ihed.ras.ru

ЗИЦЕРМАН Владимир Юрьевич – кандидат физико-математических наук, ведущий научный сотрудник лаборатории теплофизических баз данных ОИВТ РАН
e-mail: vz1941@mail.ru

КОБЗЕВ Георгий Анатольевич – доктор физико-математических наук, главный научный сотрудник лаборатории теплофизических баз данных ОИВТ РАН
e-mail: gkbz@mail.ru

КОСИНОВ Андрей Владимирович – инженер-программист лаборатории теплофизических баз данных ОИВТ РАН
e-mail: kosinov@gmail.com

УДК 004.65:001.89–047.44

И.В. Селиванова, Д.В. Косяков, А.Е. Гуськов

Влияние ошибок в базе данных *Scopus* на оценку результативности научных исследований*

На основе случайной выборки профилей 400 российских авторов и 400 организаций рассматриваются причины возникновения профилей-дублей в базе данных Scopus. Оценивается количество профилей-дублей, анализируется погрешность, которую могут вносить ошибки в библиографических описаниях в результаты наукометрических исследований, основанных на базе данных Scopus. Анализ показал, что в Scopus 76% организаций и 24% авторов имеют профили-дубли. В связи с этим организации теряют в среднем 17% публикаций, авторы – 11%. Результаты исследования могут быть использованы при корректировке базы данных Scopus и оценке погрешности при исследовании результативности научной деятельности.

Ключевые слова: *библиографические базы данных, Scopus, идентификация, наукометрия, библиометрия, библиографические ошибки, ORCID*

ВВЕДЕНИЕ

Среди целей научно-технического развития России, определенных в майских указах Президента РФ в 2012 г., было обозначено увеличение числа российских публикаций в базе данных *Web of Science* (*WoS*) до 2,44% к 2015 г., а также вхождение не менее пяти университетов России в первую сотню ведущих мировых университетов к 2020 г. В связи с этим национальная научная политика все больше ориентируется на количественные показатели публикационной активности российских ученых.

Источником информации для оценки публикационной активности служат библиографические базы данных (ББД), такие как *WoS*, *Scopus*, РИНЦ. Руководство научных организаций использует их при оценке научной результативности сотрудников и при расчете стимулирующих выплат, что мотивирует научных работников публиковать результаты исследований в журналах, индексируемых в международных базах данных.

Национальная и ведомственная оценки результативности научной деятельности организаций, выполняющих научно-исследовательские, опытно-конструкторские и технологические работы также основываются на сведениях о публикационной активности, взятых из этих баз данных [1, 2]. Наукометрические показатели, получаемые из ББД, рассматриваются как истина в последней инстанции, но при работе с ними не учитываются ошибки, влияющие на эти оценки.

Цель нашей статьи – анализ погрешности, которую могут вносить библиографические ошибки в российские исследования, связанные с оценкой результативности научной деятельности, источником данных для которых служит *Scopus*.

ОБЗОР ИССЛЕДОВАНИЙ ПО ПРОБЛЕМЕ

Качество ББД изучалось в некоторых зарубежных работах. Так в статьях F. Franceschini, D. Maisano, L. Mastrogiacomo, посвященных выявлению основных ошибок, встречающихся в *Web of Science* и *Scopus*, выделено два типа подобных ошибок [3, 4].

1. Ошибки, сделанные авторами / издателями / редакторами при подготовке списка литературы:
   - пропущенный или неверный заголовок статьи в списке литературы;
   - ошибки в других полях, таких как имя автора, название журнала, год, том и номер журнала.
2. Ошибки, возникшие в БД при связывании (т.е. установлении идентифицирующих связей между объектами в одной или разных системах, например, публикация – автор, автор – аффилиация) статей:
   - ошибка в имени автора, допущенная при переносе в БД (например, *Özel* переносится как *Oezel*);
   - расширенный список литературы (например, вместо 24-х реальных ссылок в *Scopus* указано 192, из которых 168 возникли в результате «фантомного цитирования», т. е. добавления публикаций, не упоминавшихся в оригинальной версии статьи);
   - полупустой список литературы (вместо библиографического описания указано *Reference information not available*);
   - полностью отсутствующий список литературы;
   - неверный или отсутствующий *Digital Object Identifier* (DOI);
   - потеря цитирования из-за статей вида *Online-first*;

* Работа выполнена в рамках темы научно-исследовательских работ №0334-2019-006 при поддержке Российского фонда фундаментальных исследований, грант № 18-011-00797.

- непроиндексированные статьи (иногда БД «забывают» проиндексировать некоторые статьи, хотя другие публикации из этого же выпуска журнала могут быть проиндексированы. Расчет частотности этой ошибки показал, что она характерна в большей степени для *Scopus*, чем для *WoS*);
- статья проиндексирована в БД, но по неизвестной причине отсутствует в списке литературы процитировавшей ее статьи.

Эту же классификацию ошибок применяет R.A. Buchanan в работе с использованием *Science Citation Index Expanded* и *SciFinder Scholar* [5]. Проблемы со списком литературы обсуждают в своей статье N.J. van Eck и L. Waltman [6].

Исследования показывают, что такие ошибки носят систематический характер и приводят к трудностям при поиске публикаций, а также к сильному искажению библиометрических показателей, относящихся к журналам, ученым или научным организациям. Массовое использование глобальных идентификаторов публикаций *DOI* частично решило проблему идентификации публикаций в пристатейных списках литературы при формировании индексов цитирования. Однако ситуация по-прежнему далека от идеальной: один и тот же идентификатор может встречаться у двух разных статей, одна статья может иметь два разных идентификатора, а сам идентификатор может быть указан неверно (в частности, из-за путаницы символов “O” и “0”, “O” и “Q”, “b” и “6”) [7, 8].

Рассматриваются и ошибки, связанные с появлением дублей статей, которые возникли из-за неверного отнесения статьи к журналам одного и того же издательства, а также из-за орфографических различий и изменений названия журнала [9].

В нескольких публикациях устанавливаются причины возникновения ошибок в именах зарубежных авторов. В работе C. Demetrescu и соавторов [10] разбираются ошибки в написании имен итальянских авторов в четырех базах данных: *WoS*, *Scopus*, *PubMed*, *CrossRef*; в табл. 1 представлены выделенные типы некорректного написания.

S. Ainsworth и J.M. Russell сравнивают соотношение ошибочно написанных имен испанских авторов в различных БД на примерах из латиноамериканского журнала «*Investigación Bibliotecológica*» за 2012 и 2015 гг. [11]. Ошибки в основном касаются порядка отцовских и материнских фамилий, что характерно для испанских авторов. В работе V. Aman на примере лауреатов премии Лейбница оценивается количество дубликатов *Scopus AuthorID* [12].

Общее влияние качества массива данных на проведение библиометрического анализа при составлении авторских рейтингов и регрессионном анализе, где в качестве зависимых переменных выступает число публикаций, оценивает J. Schulz [13]. Результаты этого исследования показали, что ранжирование авторов можно проводить только при очень высоком качестве данных. Хотя рассматривалось только несоответствие имен авторов, J. Schulz приходит к выводу, что ошибки в названиях организаций тоже будут оказывать большое влияние при составлении различных рейтингов (например, рейтингов университетов). Для регрессионного же анализа с наборами библиометрических данных должны проверяться систематические различия в ошибках, которые коррелируют с независимыми переменными (например, если использовать в качестве независимой переменной страну) и могут влиять на достоверность результатов регрессии.

Обобщая зарубежные исследования, можно дать следующую классификацию ошибок в библиографическом описании:

1. Список литературы
   a. Неточность в списке цитируемой литературы
   b. Отсутствие списка литературы
   c. Неверное количество источников в списке литературы
2. Статья
   a. Ошибка в названии
   b. Ошибки цитирования статей из-за публикаций online-first
   c. Ошибки в DOI
3. Журнал
   a. Опечатки в названии, годе издания, томе или номере журнала
   b. Изменение названия журнала
   c. Различные варианты написания названия одного и того же журнала
4. Авторы
   a. Составные фамилия и имя
   b. Диакритические знаки
   c. Апострофы
   d. Опечатки в фамилии и имени

Таблица 1

Типы некорректного написания имен итальянских авторов

Тип некорректного написания	Правильное написание	Некорректное написание
Часть составной фамилии переходит в имя и становится инициалом	De Rossi, Giuseppe	Rossi, G.D.
Часть составного имени становится фамилией	Verdi, Carlo Maria	Maria Verdi, C.
Одна или более частей исчезают из составной фамилии	La Torre, V	Torre, V
Из фамилии исчезают диакритические знаки	Trifrò, S.	Trifro, S.
Диакритические знаки из фамилии отображаются некорректно	Spanò, A.	Span□, A.
Из фамилии исчезает апостроф	D’Innocenzo F.	Dinnocenzo F.
Опечатки в фамилии	Accornero, F.	Accomero, F.
Опечатки в имени	Bianchi, Erica	Bianchi, Enrica

Если ошибки, связанные с журналами, статьями и списками литературы, подходят и к российскому случаю, то случаи 4.a, 4.b и 4.c применимы только к фамилиям зарубежных авторов.

У написания имен, отчеств и фамилий российских авторов и написания названий научных организаций России существует своя специфика. Она связана как с транслитерацией букв кириллического алфавита, не имеющих аналогов в латинице (например, *ё, ю, ш, щ, ъ, ы, ь*), так и с иерархией подчиненности научных организаций в России (имеется в виду принадлежность организаций к государственным академиям наук РФ, либо к их региональным отделениям), а также с ее изменениями в последние годы. Эти ошибки служат причиной неверной привязки публикаций, что в свою очередь ведет к появлению профилей-дублей авторов и организаций.

С целью однозначного определения фамилий авторов в работе [14] предложена система уникальных идентификаторов *ORCID*, в последние годы широко признанная научным сообществом и принятая многими издателями научных журналов. Этот идентификатор не завязан на владельце ББД, используется как в *Scopus*, так и в *WoS* и *Dimensions*. Таким образом, *ORCID* может быть идентификатором, претендующим на глобальное признание [15].

Многие журналы требуют указание идентификатора *ORCID* в списке авторов научной статьи. *ORCID* используется и в национальных системах научной информации [16]. Ранее компанией *Thomson Reuters* был предложен универсальный идентификатор автора *ResearcherID*, который не получил широкого признания в связи с тесной связкой с базой данных *WoS*. Компания *Elsevier* в системе *Scopus* использует уникальные идентификаторы авторов *Scopus AuthorID*. Все эти идентификаторы часто применяются в различных информационно-аналитических системах научных институтов при интеграции данных о публикационной активности их сотрудников [17]. Но при связывании данных все равно отмечается необходимость их визуальной проверки [18].

К сожалению, сами авторы плохо отслеживают свои публикации [19], что приводит к неполноте их профилей в ББД, и, соответственно, к некорректным результатам связывания при использовании систем идентификации.

Предпринимаются попытки и для введения уникальных идентификаторов организаций. Одним из таких проектов является *Research Organization Registry* (ror.org), инициированный *California Digital Library*, *Crossref*, *DataCite* и *ORCID*. Однако этот проект находится в ранней стадии, а предыдущие инициативы, такие как идентификатор организации *ISNI* [20], не были широко приняты сообществом издателей и владельцев ББД.

МЕТОДЫ ИССЛЕДОВАНИЯ

Данные

Для нашего анализа была выбрана БД *Scopus*, в которой у каждого автора и организации может быть свой профиль с идентификатором. Данные о публикациях российских авторов были получены с помощью *Scopus Search API* и *Scopus Abstract Retrieval API* в 2017 г.

Дубли организаций

Наличие профиля-дубля у российских организаций проверялось на случайной выборке из 400 профилей. Это количество было получено из формулы В.И. Паниотто с тем условием, что доверительная вероятность P=0,954. Для указанной доверительной вероятности коэффициент доверия $t = 2$, дисперсия качественного альтернативного признака принимается максимально возможной, т.е. соответствующей доле 0,5 [21, с.62]:

$$n = \frac{t^2 p(1-p)}{\Delta^2} = \frac{2^2 \cdot 0{,}5(1-0{,}5)}{\Delta^2} = \frac{1}{\Delta^2} = \frac{1}{0{,}05^2} = 400 , \quad (1)$$

где n – объем выборки; p – генеральная доля единиц, обладающих значением признака, относительно которого рассчитывается ошибка выборочной доли; t – доверительный коэффициент; Δ – предельная ошибка выборки.

Для того чтобы анализ был более точным, введено ограничение на количество публикаций в профиле организации: в 2017 г. их должно быть не менее 5. При меньшем количестве публикаций в список попадали организации, реальное название которых сложно было определить. Например, организация с *Affiliation ID*=119911371 называлась *Berdsk*, получившееся из-за перехода в название части адреса. Также были обнаружены явные дубли других организаций. В качестве примера приведем организацию с *Affiliation ID*= 109886056, в названии которой указано *Romsk Polytechnical University*. Можно предположить, что этот профиль, скорее всего, является дублем профиля Томского политехнического университета (ТПУ). Подтвердить это предположение невозможно, так как в этом профиле только одна публикация, в аффилиациях авторов которой указана только эта организация. В основном профиле ТПУ этих авторов обнаружить не удалось.

Алгоритм анализа профилей организаций

1. Для организации формируем список из 10 аффилированных с ней авторов с наибольшим числом публикаций: $A_i = \{a_1, a_2, \ldots, a_M\}$, где M = 10.
2. Для каждого автора a_i, i=1,2, …, M формируем полный список его аффилиаций $OA_i = \{oa_{i1}, oa_{i2}, \ldots\}$.
3. Формируем общий список аффилиаций 10 самых продуктивных авторов из организации $GL = \bigcup OAi$.
4. В списке *GL* выявляем дубликаты – переводы названий, сокращенные и схожие названия.

Алгоритм применялся для каждой из 400 организаций и позволил выявить дублирующие профили аффилиаций. При сопоставлении названий дополнительно изучались сайт организации, «Википедия», а также профиль организации в Научной электронной библиотеке eLibrary.ru.

Дубли авторов

Наличие профиля-дубля у автора проверялось по формуле (1) на случайной выборке из 400 активных российских авторов (не менее двух публикаций в 2017 г.). Ограничение на количество публикаций бы-

ло введено из-за того, что при отсутствии этого фильтра в случайную выборку попадало слишком много профилей авторов с одной публикацией, что не позволяло проанализировать наличие у этих авторов дублирующих профилей. В списках российских авторов из Scopus было обнаружено около 47% таких профилей.

Алгоритм поиска профиля-дубля автора

Обозначим множество авторов за $A = \{a_1, a_2, \ldots, a_N\}$, где $N = 400$.

1. Для каждого автора a_i, i=1, …, N формируем полный список аффилиаций $OA_i = \{oa_{i1}, oa_{i2}, \ldots\}$, т. е. для каждого из авторов – полный список организаций, о которых есть упоминание в его публикациях.
2. Для каждой организации oa_{ij} формируем полный список авторов $C_{ij} = \{c_{ij}^1, c_{ij}^2, \ldots\}$, т. е. авторов, работающих в одной организации o_{ij} с автором a_i, включая и самого автора a_i.
3. Формируем общий список коллег автора a_i из всех организаций $C_i = \bigcup_j C_{ij}$.
4. Из списка C_i находим все дубликаты – авторы с ФИО автора a_i, которое либо полностью совпадает с его ФИО, либо похоже на него. Анализ проводился вручную, и совпадения проверялись экспертным путем (при автоматизированном поиске необходимо использовать различные меры близости).

Для того чтобы исключить возможность определения профиля полного тезки автора как профиля-дубля исследуемого автора, дополнительно анализировались списки соавторов, тематики исследования, организаций, полные тексты статей, списки авторов на eLibrary.ru и аффилиационная история. Мы полагаем, что если у нескольких профилей с одинаковыми ФИО совпадают места работы, соавторы и тематика, то это профили одного и того же автора.

Этапы исследования

Каждый из трех этапов исследования выполнялся как для российских авторов, так и для организаций.

1. Определение причины возникновения профилей-дублей
2. Оценка количества таких профилей
3. Оценка доли публикаций в профилях-дублях от общего количества.

На каждом этапе результаты были получены и для организаций, и для авторов.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Оценка количества и определение причин возникновения профилей-дублей организаций

Анализ профилей организаций показал следующие результаты.

A. 19% (75 из 400) организаций дублей не имеет.

B. 5% (20 из 400) профилей организаций содержат ошибки (иностранные организации, разные адреса и др.), что затрудняет определение дублей.

C. 76% (305 из 400) организаций имеют хотя бы один дубль.

Рассмотрим подробнее ошибки, обнаруженные в профилях 5% организаций (результат ***B***).

Случай B1. *Вместо названия организации указаны профили: РАН, СО РАН и др.*

Ошибка возникает из-за неверной привязки организации, в названии которой встречается, например, Сибирское отделение Российской академии наук (СО РАН). Такая организация может быть отнесена к профилю как СО РАН, так и РАН, при этом в последнем случае публикация будет приписана Москве, а не Новосибирску или другому сибирскому научному центру. Это может повлиять на результаты некоторых исследований, например, по академической мобильности ученых (*Affiliation ID* профиля СО РАН в *Scopus* – 60017604, РАН – 60021331).

Для оценки количества публикаций, которые может потерять организация из-за отнесения к профилям РАН, РАМН или их региональным отделениям, была сделана случайная выборка 400 профилей российских авторов и посчитана доля публикаций для случая B1. Результаты показали, что из-за неверного отнесения публикаций авторов к государственным академиям наук РФ в среднем для одного автора организация теряет до 14% публикаций.

Случаи, когда автор действительно может указывать аффилиацию Российской академии наук или ее региональных отделений, являются редкими.

Случай B2. *Иностранные организации*

Ошибка возникает из-за неверного распознавания названия города. Например, университет Айдахо расположен в городе *Moscow*, штат Айдахо, США. Но в *Scopus* один из профилей университета относится к России (*Affiliation ID*=100804052, количество публикаций = 106). При этом в список из 400 организаций попал еще один профиль университета Айдахо, также отнесенный к российским организациям (*Affiliation ID*=110070192, количество публикаций = 73).

Случай B3. *Сложносоставные организации с разными подразделениями*

К данному случаю относятся как университеты, где авторы иногда указывают только название факультета, и такая публикация может попасть в другой профиль, так и организации, подразделения которых находятся в разных городах. В качестве примера приведем организацию с *Affiliation ID*=100465640. В ее названии указано *Chemistry Department*. При этом в адрес организации попало полное название *Lomonosov MSU*, *Moscow*.

В табл. 2 указаны причины возникновения дублей профилей российских организаций (результат C), а также частота их встречаемости на примере тех профилей, которые имеют только один дубль (общим количеством 66).

Данные в ББД *Scopus* активно обновляются и их качество растет, как в связи с усилиями специалистов *Elsevier*, так и с обращениями пользователей системы. Результаты нашего исследования показали, что из 66 дублирующихся профилей на начало 2017 г. почти 41% оказались объединенными к марту 2019 г.

Таблица 2

Анализ профилей с одним дублем

Случай	Причина возникновения дублей	Кол-во профилей	Доля в общем количестве, %
C1	Старое название организации	1	2
C2	Сокращенное название организации, включая • частичную аббревиацию названия (например, вместо *the Siberian Branch of the RAS* указано *SB RAS*); • сокращенное название, добавленное к полному	13	20
C3	Отсутствие части названия	17	26
C4	Отсутствие имени/части имени ученого, присвоенного организации, его инициалов	3	5
C5	Попадание части названия организации в адрес	2	3
C6	Неверный/неполный адрес организации (включая неверное указание страны или города)	4	6
C7	Разный порядок слов в названии	1	2
C8	Полное совпадение названий (например, у *Tver State Medical University* имелся дубль с таким же названием)	4	6
C9	Разная транслитерация названия	2	3
C10	Разный перевод названия (включая перевод не только на английский язык)	9	14
C11	Ошибки в названии (опечатки, лишние пробелы и знаки препинания)	10	15
C12	Зависимость от регистра	0	0

Таблица 3

Профили-дубли Национального медицинского исследовательского центра психиатрии и наркологии имени В.П. Сербского

Affiliation ID	Название организации	Ошибка
60085185	Serbsky Institute for General and Forensic Psychiatry	C1
100312210	Serbsky Res. Inst. Forens. Psychiat.	C1, C2
100331806	V. P. Serbskii Central Research Institute of Forensic Psychiatry	C4, C9, C11
100360057	Serbsky National Research Center for Social and Forensic Psychiatry	C1
100366550	Serbsky Natl. Res. Ctr. Social F.	C1, C2
100368625	V. P. Serbskii State Research Center	C9, C10
100390434	Gos. Nauchnyj Tsentr Sotsial'noj	C3
100755591	NI Inst. Obshchej/Sudebnoj Psikhiat.	C1, C2, C4, C9
101328961	Serbskii State Scientific Center for Social and Forensic Psychiatry	C1, C9
101335050	Lab. of Fed. State Inst. State Sci. Centre of Social and Forensic Psychiatry after V.P. Serbsky	C1, C7
101981345	V. P. Serbskii State Research Center of Social and Forensic Psychiatry	C1, C4, C9
105353230	Serbsky National Research Centre for Social and Forensic Psychiatry	C1
108509206	Serbsky's Institute of General and Forensic Psychiatry	C1, C10
108847913	Serbsky National Research Centre for Social and Forensic Psychiatry	C1
112568830	V. P. Serbskii State Scientific Center for Social and Forensic Psychiatry	C1, C4, C9, C11
112615739	Serbsky State Research Center of Social and Forensic Psychiatry	C1
112617371	V. P. Serbsky State Research Center for Social and Forensic Psychiatry	C1, C11
112625607	V. P. Serbskii Center of Social and Forensic Psychiatry	C1, C4, C9, C11
112662194*	V. P. Serbskii State Research Center of Forensic	C1, C3, C4
112956380	V.P. Serbsky State Scientific Center for Social and Forensic Psychiatry	C1, C4
113068622	Serbsky National Research Center for Social and Forensic Psychiatry	C1
113157802	Serbsky State Scientific Center for Social and Forensic Psychiatry	C1
113939370	Serbsky National Research Center	C3
114190516	Serbsky State Scientific Center of Social and Forensic Psychiatry	C1
115403520	Serbsky Federal Medical Research Center of Psychiatry and Narcology	C10
116873944	Serbsky Federal Medical Research Center for Psychiatry and Narcology	C10

* профиль, отсутствующий в *Scopus* на 04.05.19

Максимальное количество – 26 дублей обнаружено в *Scopus* в названии Федерального государственного бюджетного учреждения «Национальный медицинский исследовательский центр психиатрии и наркологии имени В.П. Сербского» Министерства здравоохранения Российской Федерации, сокращённо – ФГБУ «НМИЦПН им. В.П. Сербского» Минздрава России, старые названия которого: «Государственный научный центр социальной и судебной психиатрии им. В.П. Сербского», «Федеральный медицинский исследовательский центр психиатрии и наркологии имени В. П. Сербского. Все найденные профили этой организации представлены в табл. 3. Основным считаем профиль, имеющий максимальное число публикаций.

Оценка количества и определение причин возникновения профилей-дублей авторов

Анализ профилей авторов показал следующие результаты.

A. У 75% (300 из 400) авторов не было обнаружено профилей-дублей.

B. У 1% (3 из 400) не удалось установить наличие дубля.

Это связано, например, с тем, что в профиле организации в *Scopus* встречается несколько авторов с одним и теми же фамилией и инициалами, но в профиле организации, как и на сайте института, такого сотрудника нет. В профилях этих авторов не совпадают также и соавторы. В другом случае имя и фамилия автора имеют вид Alexander R., т. е. определить фамилию этого автора становится невозможным.

C. У 24% (97 из 400) были обнаружены профили-дубли.

Необходимо отметить, что с начала 2017 г. к марту 2019 г. 9% (36 из 400) профилей авторов объединены в *Scopus* полностью, 2% (7 из 400) – частично, а 13,5% (54 из 400) профилей так и остаются необъединенными.

Рассмотрим причины и частоту возникновения 61 профиля-дубля из тех, которые полностью (54 профиля) или частично (7 профилей) не объединены в *Scopus*.

Случай 1. *Разная транслитерация (41 из 61, 67%)*

Разная транслитерация фамилии, имени или отчества автора появляется из-за наличия в них букв *ё, ю, я, х, ц, ш, ь* и др. (например, имя одного автора транслитерируется в одном случае как Yury, в другом – Yrii).

Примером автора, в дублирующих профилях которого наблюдается разная транслитерация и в имени, и в фамилии, является Bakaeva Natalia Vladimirovna (*AuthorID* = 57195920506). В ее профиле-дубле (*AuthorID* = 56826095700) имя указано как Bakayeva, Natalya. Для этого примера отметим еще один важный момент: обоим профилям дается один и тот же *ORCID*, но профили не объединены, т. е. при объединении профилей в *Scopus* механизм объединения по совпадающему *ORCID* либо отсутствует, либо срабатывает не всегда.

Случай 2. *Ошибка в фамилии (10 из 61, 16%).*

- ***Лишние буквы, опечатки, ошибки распознавания, похожие по внешнему виду буквы (например, заглавная I и строчная l), строчные и заглавные буквы, ошибки в самой публикации.***

Одной из иллюстраций этого типа ошибок является профиль автора (основной *AuthorID* = 7202813119) Костромского государственного университета – Smirnova, N.A. Второй ее профиль (*AuthorID* = 56568452600) обозначен как Smimova, N. A., т.е. буквы rn в фамилии были распознаны как m. В ее третьем профиле (*AuthorID* = 55480627000) в качестве ФИО указано: SMlRNOVA, N. A., т.е. буква I была распознана как l.

В другом примере ошибочное написание фамилии автора возникло уже в исходной публикации. Среди профилей авторов Научно-исследовательского института фармакологии им. В.В. Закусова были обнаружены два автора: Seredenin, Sergey B. (*AuthorID* = 7004680975) и Seredin, S. B. (*AuthorID* = 6504406983). При этом в списке авторов на сайте eLibrary.ru автора с фамилией Середин обнаружено не было. В полном тексте статьи из профиля автора Seredin, S. B. также была указана фамилия Середин. Через eLibrary.ru было выяснено, что статья является переводной версией русскоязычной публикации, в которой автором указан Середенин С.Б. О том, что эта статья принадлежит именно Середенину, также свидетельствует наличие общих соавторов с основным профилем, список литературы, где практические все работы принадлежат Середенину.

- ***К окончанию фамилии приписана буква от аффилиации.***

Примером может послужить профиль автора из Института общей и неорганической химии имени Н.С. Курнакова РАН – Fedorchenko, Irina Valentinovna (*AuthorID* = 24474356600). У нее есть профиль-дубль с одной публикацией, где в имени профиля написано Fedorchenkob, I.V. Детальный анализ полного текста этой публикации показал, что *b* в конце фамилии появилась из-за обозначения этой буквой ссылки на организацию автора; в полном тексте публикации данной ошибки нет.

Случай 3. *Разный вариант представления имени (8 из 61, 13%).*

- ***Вместо полного имени/отчества указаны инициалы; отчество отсутствует полностью.***

Рассмотрим профиль автора из Института общей физики имени А.М. Прохорова РАН Kozlov, D.N. (*AuthorID* = 55407907300). В профиле у автора 75 публикаций. Для этого ученого был найден еще один профиль: Kozlov, Dimitry N. (*AuthorID* = 26650432300) с одной публикацией. Соавторы и тематики исследований в профилях совпадают.

- ***Перепутаны местами имя, отчество и фамилия.***

У автора из Университета ИТМО – Leonov Mikhail Yu – два профиля: в первом (*AuthorID* = 36604774600) в качестве фамилии указано: Leonov, имени: Mikhail Yu; во втором (*AuthorID* = 57192376476) фамилия у автора уже Leonov Mikhail, в имени указаны лишь инициал отчества Yu. В основном профиле у автора 39 публикаций, в дубле – одна.

Другим подобным примером является профиль автора из Института гидродинамики имени М.А. Лаврентьева СО РАН, в основном профиле которого (*AuthorID* = 6603685354) указаны фамилия: Liapidevskii, имя: Valery Yu. В одном из его профилей-дублей (*AuthorID* = 57196437043) в фамилию также вошел инициал от отчества: Yu Liapidevskii, а имя осталось верным: V.

Случай 4. *Разные периоды публикаций (1 из 61, 2%).*

Единственным примером, относящимся к этому случаю, является профиль автора из Рязанского государственного медицинского университета. В одном профиле (*AuthorID* = 6505788719) охват публикаций с 1985 по 1993 гг., в другом (*AuthorID* = 57195313658) – с 2017 по 2018 гг. Профиль автора был найден также в eLibrary.ru, где охват публикаций с 1986 по 2018 гг. С профилем в eLibrary.ru совпадают списки статей и большинство соавторов как из первого, так и из второго профиля. Стоит отметить, что на eLibrary.ru возможны, хотя и редки, случаи неверного объединения профилей авторов, но в этом случае из-за совпадения большинства соавторов профиль объединен верно. Дополнительного третьего профиля в *Scopus*, в котором могли бы находиться публикации автора за неотраженный период (с 1994 по 2016 гг.), обнаружить не удалось.

Оценка доли публикаций в профилях-дублях от общего количества

На третьем этапе исследования для оценки количества публикаций, которые могут пропасть из основного профиля, был введен показатель $L = \frac{S(D)}{S(O+D)}$, где $S(D)$ – суммарное число публикаций во всех профилях-дублях, а $S(O+D)$ –общее число публикаций в основном профиле и его профилях-дублях.

Анализ показал, что для организации средние потери публикаций могут составлять 17% (максимальное значение – 83% достигается для уже упомянутого *Serbsky Federal Medical Research Centre for Psychiatry and Narcology*, *Affiliation ID* основного профиля = 116873944), для авторов – 11% (максимальное значение – 55%, *AuthorID* основного профиля = 8630661000).

ЗАКЛЮЧЕНИЕ

Проведенное нами исследование выявило, что 76% организаций и 24% авторов научных публикаций имеют профили-дубли в *Scopus*. Дубли профилей организаций возникают по следующим причинам: различия при транслитерации и переводе названий организаций, опечатки и ошибки в названиях, отсутствие части названия, указание разных адресов, неверное указание принадлежности к тому или иному региональному отделению Российской академии наук.

Дубли профилей авторов также возникают из-за различий при транслитерации и опечаток. Характерной причиной является ошибочное написание фамилии, имени или отчества: к фамилии приписывается символ, взятый из обозначения организации или из отчества; фамилия и имя могут быть перепутаны местами. Были обнаружены и случаи, когда вместо фамилии указано сокращение, что не позволяло идентифицировать автора.

Из-за подобных ошибок в *Scopus* искажается количественная оценка результативности: организации в среднем теряют 17% публикаций, а авторы – 11%.

Наш анализ дает дополнительную информацию для корректировки ошибок в профилях авторов и организаций, которая постоянно происходит в *Scopus*. Вместе с тем, при наукометрическом анализе оценке результативности научной деятельности на основе данных *Scopus* необходимо учитывать погрешность, вызванную некорректной идентификацией и связыванием авторов, организаций и публикаций.

СПИСОК ЛИТЕРАТУРЫ

1. Гуськов А.Е., Косяков Д.В., Селиванова И.В. Методика оценки результативности научных организаций // Вестник Российской академии наук. – 2018. – Т. 88, № 5. – С. 430-443.
2. Kosyakov D., Guskov A. Research assessment and evaluation in Russian fundamental science // Procedia Computer Science. – 2019. – Vol. 146. – P. 11-19
3. Franceschini F., Maisano D., Mastrogiacomo L. Empirical analysis and classification of database errors in Scopus and Web of Science // Journal of Informetrics. – 2016. – Vol. 10, Issue 4. – P. 933-953.
4. Franceschini F., Maisano D., Mastrogiacomo L. The museum of errors/horrors in Scopus // Journal of Informetrics. – 2016. – Vol. 10, Issue 1. – P. 174-182.
5. Buchanan R.A. Accuracy of cited references: The role of citation databases// College & Research Libraries. – 2006. – Vol. 67, Issue 4. – P. 292-303.
6. Nees Jan van Eck, Ludo Waltman. Accuracy of citation data in Web of Science and Scopus // arXiv:1906.07011. – URL: https://arxiv.org/ftp/arxiv/papers/1906/1906.07011.pdf
7. Franceschini F., Maisano D., Mastrogiacomo L. Errors in DOI indexing by bibliometric databases // Scientometrics. – 2015. – Vol. 102, Issue 3. – P. 2181–2186.
8. Zhu J., Hu G., Liu W. DOI errors and possible solutions for Web of Science // Scientometrics. – 2019. – Vol. 118, Issue 2. – P. 709–718.
9. Valderrama-Zurián J.-C., Aguilar-Moya R., Melero-Fuentes D., Aleixandre-Benavent R. A systematic analysis of duplicate records in Scopus// Journal of Informetrics. – 2015. – Vol. 9, Issue 3. – P. 570-576.
10. Demetrescu C., Ribichini A., Schaerf M. Accuracy of author names in bibliographic data sources: an Italian case study // Scientometrics. – 2018. – Vol. 117, Issue 3. – P. 1777–1791.
11. Ainsworth S., Russell J.M. Has hosting on science direct improved the visibility of Latin American scholarly journals? A preliminary analysis of data quality // Scientometrics. – 2018. – Vol. 115, Issue 3. – P. 1463–1484.
12. Aman V. Does the Scopus author ID suffice to track scientific international mobility? A case study

based on Leibniz laureates // Scientometrics. – 2018. – Vol. 117, Issue 2. – P. 705-720.
13. Schulz J. Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analysis // Scientometrics. – 2016. – Vol. 107, Issue 3. – P. 1283–1298.
14. Haak L.L., Fenner M., Paglione L., Pentz E., Ratner H. ORCID: A system to uniquely identify researchers // Learned Publishing. – 2012. – Vol. 25, Issue 4. – P. 259-264.
15. Mazov N.A., Gureyev V.N. Modern challenges in bibliographic metadata identification. 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), Vladivostok // IEEE. – 2018. – P. 1-4.
16. Moreira J.M., Cunha A., Macedo N. An ORCID based synchronization framework for a national CRIS ecosystem // F1000Research. – 2015. – Vol. 4. – P. 181.
17. 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. – Т. 9. – С. 95-102.
18. Ковязина Е.В. Корпоративные репозитории научных публикаций и проблемы обмена данными // Труды ГПНТБ СО РАН. – 2016. – Т. 10. – С. 288-292.
19. Захарова С.С., Гуреева Ю.А. Научные публикации: от картотеки трудов до библиографических профилей // Библиосфера. – 2017. – №2. – С.85-89
20. MacEwan A., Angjeli A., Gatenby J. The international standard name identifier (ISNI): The evolving future of name authority control // Cataloging and Classification Quarterly. –2013. – Vol. 51, Issue (1-3). – P. 55-71.
21. Могильчак Е.Л. Выборочный метод в эмпирическом социологическом исследовании: учеб. пособие. – Екатеринбург: Изд-во Уральского ун-та, 2015. – 120 с.

Материал поступил в редакцию 05.07.19

Сведения об авторах

СЕЛИВАНОВА Ирина Вячеславовна – младший научный сотрудник ГПНТБ СО РАН
e-mail: selivanova@spsl.nsc.ru;

КОСЯКОВ Денис Викторович – зам. директора по развитию ГПНТБ СО РАН
e-mail: kosyakov@spsl.nsc.ru

ГУСЬКОВ Андрей Евгеньевич – кандидат технических наук, директор ГПНТБ СО РАН
e-mail: guskov@spsl.nsc.ru

СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ

УДК 021:002.6:006.72

А.А. Джиго, Т.В. Майстрович

Новый уровень понимания библиотечно-информационной услуги*

Раскрывается содержание понятия «библиотечно-информационная услуга»; рассматривается минимальный перечень услуг по библиотечному, библиографическому и информационному обслуживанию в научно-информационных учреждениях различного уровня (центральные, отраслевые, локальные) и показатели, позволяющие судить об эффективности оказания этих услуг в соответствии с требованиями ГОСТ Р 7.0.104-2019.

Ключевые слова: *научная библиотека, библиотечно-информационная услуга, режимы и опции, стандартизация*

Вкладом в совершенствование системы отраслевых стандартов в области библиотечного дела и библиографии стала разработка и принятие 18 января 2019 г. нового национального стандарта ГОСТ Р 7.0.104-2019 «СИБИД. Библиотечно-информационные услуги научной библиотеки. Виды, формы и режимы предоставления». Этот документ разработан впервые в соответствии с «Программой национальной стандартизации на 2016–2018 гг.», одобренной Техническим комитетом ТК–191 «Система стандартов по информации, библиотечному и издательскому делу» (СИБИД) и утверждённой Федеральным агентством по техническому регулированию и метрологии Росстандарта [1], входит в систему «Стандарты по информации, библиотечному и издательскому делу» (СИБИД), научно, методически и функционально связан с другими межгосударственными и национальными стандартами [2]. Он соответствует действующим нормативно-техническим документам в области стандартизации и основывается на научно-исследовательских разработках, проведённых Фундаментальной библиотекой ИНИОН РАН, Библиотекой по естественный наукам РАН, Государственной публичной исторической библиотекой России, Центральной научной сельскохозяйственной библиотекой РАН и Всероссийским институтом научной и технической информации РАН.

В системе СИБИД есть несколько стандартов, в которых в определенном аспекте рассматриваются библиотечные услуги. В одном из аспектов, а именно статистическом, услуги впервые были отражены в ГОСТ 7.41-82 «Единицы учета обслуживания читателей и абонентов библиотек и органов научно-технической информации» [3], в другом – терминологическом – в ГОСТ 7.0-99 «Информационно-библиотечная деятельность, библиография. Термины и определения» [4]. В межгосударственном стандарте ГОСТ 7.20-2000 «Библиотечная статистика» библиотечные услуги нашли более глубокую проработку. Большая детализация такого важного компонента деятельности библиотек и информационных служб как библиотечная услуга дана в национальных стандартах: ГОСТ Р 7.0.20–2014 «Библиотечная статистика: показатели и единицы исчисления» (статистические показатели) [5] и ГОСТ Р 7.0.103–2018 «Библиотечно-информационное обслуживание. Термины и определения» (терминология, относящаяся к библиотечной услуге) [6]. Однако обобщенного специализированного технологического стандарта до 2018 г. в российской практике не было. Поэтому ГОСТ Р 7.0.104-2019 разработан и принят впервые.

Концепция стандарта, все промежуточные версии открыто обсуждались с экспертами и библиотечными специалистами. Первая редакция этого документа была разослана в 120 библиотек и органов научно-технической информации Российской Федерации. Получены отзывы от 97 организаций, что нашло отражение в сводке отзывов – 134 замечания. Уведомление о разработке стандарта было размещено на сайте Федерального агентства по техническому регулированию и метрологии 26 июля 2018 г, уведомление о завершении публичного обсуждения проекта

* К принятию ГОСТ Р 7.0.104-2019 «СИБИД. Библиотечно-информационные услуги научной библиотеки. Виды, формы и режимы предоставления»

также – 26 октября 2018 г. На основании предложений и замечаний подготовлена вторая редакция национального стандарта, проведено согласительное совещание. Окончательный текст согласован с 12 библиотечно-информационными организациями [1].

Известно, что разработка стандарта проходит в несколько этапов. Рассматриваемый стандарт не исключение. На первом этапе были созданы базовые теоретико-методические положения стандарта и утверждена его концепция.

Поскольку национальный стандарт не может отражать чью-либо частную точку зрения, то по отдельным вопросам находились компромиссные решения, которые уточнялись по результатам экспертных обсуждений. Научно-методологические основы проекта прошли многократное обсуждение на специализированных семинарах и заседаниях рабочей группы (общее количество проведенных профильных мероприятий – 14). Одно из наиболее значимых – это заседание Секции 31 по научно-исследовательской работе Российской библиотечной ассоциации (г. Владимир, 15 мая 2018 г.), где присутствовали авторитетнейшие специалисты отечественного библиотековедения из различных библиотек и вузов России [7].

Цель разработки стандарта – определение номенклатуры библиотечно-информационных услуг, оказываемых пользователям научных библиотек и органов НТИ. Основные задачи – раскрытие содержания библиотечно-информационных услуг; определение минимального видового перечня библиотечно-информационных услуг, необходимых пользователям научных библиотек и органов научно-технической информации; раскрытие их содержания, форм предоставления и режимов «доставки». Однако это не означает, что данный стандарт не может использоваться в деятельности библиотек других типов.

Главное, с чем согласны все рецензенты, это то, что стандарт применим для использования не только в научных библиотеках, вне зависимости от их ведомственной принадлежности и юридического статуса (самостоятельное юридическое лицо, структурное подразделение юридического лица), но и в информационных центрах и в других учреждениях для обеспечения научно- исследовательской деятельности. Он может быть использован библиотеками высших учебных заведений, которые в последнее время все чаще получают статус национальных исследовательских университетов или имеют в своей структуре научно-исследовательские подразделения. Кроме того, тщательность проработки этого документа позволяет судить о его полезности для центральных библиотек регионов и библиотек других типов.

Тем не менее, в стандарте максимально учтены виды библиотечно-информационных услуг, формы и режимы их предоставления пользователям научных библиотек и органов научно-технической информации, т. е. в основу классификации библиотечно-информационных услуг положен кластер, содержащий их важнейшие характеристики: содержание, формы представления, режимы оказания и получения, каналы доставки.

В ходе дискуссий стало очевидно, что первоначальное название стандарта «СИБИД. Библиотечно-информационная услуга. Требования для научных библиотек» неудачно. По мнению специалистов, такие показатели, как виды и формы и предоставления услуг должны найти отражение в названии документа. Соответственно в ходе работы было изменено его название.

Принятый стандарт относится к разряду технологических. Это определяет его структуру:

- область применения;
- нормативные ссылки;
- термины и определения;
- виды и формы библиотечно-информационных услуг;
- режимы предоставления и получения библиотечно-информационных услуг в научной библиотеке.

В рассматриваемом документе термины и определения приведены по ГОСТ 7.0.103 и ГОСТ 7.0 [4, 6]. Однако возникла необходимость во введении и обосновании некоторых терминов, не встречавшихся ранее в системе СИБИД:

библиометрическая услуга – предоставление данных, полученных на основе изучения профильного потока публикаций посредством библиометрического анализа;

информационный продукт – результат создания или переработки информации в целях ее многократного использования в процессе предоставления библиотечно-информационных услуг;

каналы предоставления библиотечно-информационной услуги – технико-технологические способы доставки библиотечно-информационной услуги пользователям;

номенклатура библиотечно-информационных услуг – упорядоченный перечень библиотечно-информационных услуг, предоставляемых библиотекой пользователям;

режим оказания библиотечно-информационной услуги – совокупность временных, дистанционных, технических и технологических факторов, определяющих способы и каналы доставки/предоставления конкретной формы услуги пользователю.

В разделе «Общие положения» зафиксированы новые базовые подходы к сущности библиотечно-информационных услуг. Объектом стандартизации стала библиотечно-информационная услуга, оказываемая научной библиотекой в соответствии со своими целями, задачами, запросами и потребностями пользователей. Настоящий стандарт не регулирует предоставление библиотекой технических и офисных услуг. Это связано с принятым определением библиотечно-информационной услуги как результата библиотечно-информационного обслуживания. Услуги предоставляются научной библиотекой в соответствии с ее задачами, изложенными в Уставе или Положении о деятельности. Библиотека может оказывать более широкий спектр содержательных услуг с их лицензированием в предусмотренных законодательством случаях. Другие услуги, в том числе офисные и технические не относятся к библиотечно-информационным, их предоставление или не предоставление определяется по ситуации. Помимо этого, любое учреждение оказывает такие услуги, как пре-

доставление гардероба, чистые полы, беспроводной Интернет и многое другое, но они не относятся напрямую к функционированию этого учреждения.

Объем оказываемых услуг, формы их реализации, режимы доставки и другие параметры определяются запросами пользователей научной библиотеки, направлениями ее деятельности и структурой научного учреждения.

Библиотечно-информационные услуги оказываются на базе созданных информационных продуктов, в составе которых – документные фонды, каталоги, базы данных, информационные и библиографические издания. Само производство информационного продукта рассматривается в стандарте в качестве услуги только в случае его создания в ответ на запрос пользователя.

Основанием для предоставления библиотечно-информационной услуги могут считаться индивидуальный запрос пользователей, повторяющиеся запросы, типовые запросы, запросы по номенклатуре библиотечно-информационных услуг, распоряжения руководителя научно-исследовательского учреждения. Библиотечно-информационные услуги предоставляются с использованием различной ресурсной базы: фонд научной библиотеки, фонды других библиотек; информационные продукты собственной генерации, иные информационные продукты и ресурсы.

Научные библиотеки предоставляют пользователям пять видов библиотечно-информационных услуг: библиографические, библиометрические, библиотечные, информационные, консультационные. Каждый из видов этих услуг реализуется в той или иной форме, под которой в стандарте понимается способ их предоставления в рамках существующей в научной библиотеке организации библиотечно-информационного обслуживания.

Библиографические услуги реализуются в следующих формах:

- подготовка сообщения, содержащего справку или библиографическую консультацию по запросу пользователя (детальная классификация справок приведена в ГОСТ Р 7.0.20 [5]);
- составление библиографических списков публикаций сотрудников, включая списки публикаций отдельных лиц и коллективов, а также списки публикаций, где имеются ссылки на труды заданных лиц и коллективов;
- создание библиографической продукции по индивидуальным или групповым запросам пользователей;
- повышение библиографической (библиотечно-информационной) грамотности и обучение пользователей (проведение занятий и консультаций по созданию библиографической записи, формированию библиографического аппарата научных и учебных работ – пристатейной и/или прикнижной библиографии).

Библиометрические услуги составляют специфическую особенность научных библиотек и органов НТИ и реализуются в различных формах:

- предоставление пользователю формализованных показателей и рейтингов, принятых для оценки качества научной работы;
- создание аналитического продукта на основе библиометрических и наукометрических исследований научного направления с использованием наукометрических аналитических систем;
- индивидуальный мониторинг рейтингов цитируемости ученых;
- проверка подготовленных или полученных научных работ на наличие некорректного заимствования;
- проведение поиска библиометрической информации в специализированных базах данных.

К библиотекам всех типов и видов относятся следующие библиотечные услуги и формы их оказания:

- предоставление конкретных документов во временное пользование из фонда библиотеки, фондов других библиотек;
- организация доступа (в том числе дистанционного) к полнотекстовым электронным ресурсам собственной генерации и получаемым на основе лицензионных соглашений;
- выдача текста документа (принадлежащего конкретной научной библиотеке, иной библиотеке или находящегося в составе электронного ресурса) в постоянное пользование путем его легитимного копирования;
- ознакомление пользователя с документом с помощью выставок (в том числе виртуальных), обзоров и иных форм, применяемых в библиотечно-информационном обслуживании.

Информационные услуги во многом сопрягаются с библиографическими, их разделение – это дискуссионный вопрос. После цикла обсуждений в профессиональной среде информационные услуги были представлены отдельно в таких формах, как предоставление:

- информационных продуктов любого объема, содержащих библиографическую, фактографическую информацию или их комбинацию, подготовленных научной библиотекой или полученных из других источников в постоянное или временное пользование;
- полнотекстовой информации, отобранной и систематизированной в соответствии с критериями, сформулированными в запросе пользователя.

Сущностными для библиотеки признаны следующие формы консультационных услуг:

- по пользованию библиотекой, ее справочно-поисковым аппаратом, электронными ресурсами различной генерации;
- по оформлению научных работ;
- по нормативным и организационно-распорядительным документам, необходимым для научной, образовательной, научно-организационной деятельности;
- по представлению результатов собственной научной деятельности – подаче заявок на патентование, депонированию.

Как и любые объекты, библиотечно-информационные услуги могут быть сгруппированы по различным критериям, основными из которых признаны:

- периодичность оказания: разовые (по запросу) и постоянные (по заказу или договору);

- причина оказания: по запросу пользователя, без запроса (инициативные со стороны библиотеки);
- адресность: индивидуальные (для конкретного потребителя), групповые, коллективные, массовые;
- экономические характеристики: бесплатные для конечного пользователя, платные;
- место оказания: стационарные, внестационарные.

Очевидно, что любая библиотечно-информационная услуга может получать дополнительные свойства, повышающие ее комфортность для пользователя. Дополнительные свойства могут касаться формы предоставления услуги, сроков оказания, способов доставки пользователю (список не исчерпывающий). Стандартом указывается, что дополнительные свойства услуги не являются основанием для перевода её в платную категорию, что крайне важно для пользователей.

Важно учитывать тот факт, что каждый вид библиотечно-информационных услуг оказывается в конкретных формах, но они должны быть «доставлены» потребителю тем или иным способом, который в стандарте определен как режим предоставления и получения услуги. Под режимом предоставления (получения) услуги разработчиками стандарта понимается совокупность временных, дистанционных, технических и технологических факторов, определяющих способы и каналы доставки/предоставления конкретной формы услуги пользователю. Практически любая библиотечно-информационная услуга может предоставляться в различном режиме, который не влияет на ее сущность, но может скорректировать ее востребованность.

Предоставление и получение библиотечно-информационных услуг происходит в следующих режимах:

- по степени самостоятельности работы пользователя – режим самообслуживания и услуга, предоставляемая персоналом научной библиотеки; при этом в режиме самообслуживания осуществляются отбор документов в открытом доступе, использование электронной библиотеки, иных электронных ресурсов и баз данных, в том числе электронных каталогов, заказ документов из хранилищ, копирование текстов. К услугам, оказываемым персоналом научной библиотеки, относят выдачу документов (заказ документов по МБА как промежуточный этап) или их копий;
- по синхронизации заказа, выполнения и получения услуги – режим реального времени, режим отложенного обслуживания, упреждающий режим;
- по периодичности обслуживания: разовый (режим запрос-ответ), постоянный; продолжающийся (передача информации по мере ее накопления); текущий (обслуживание с определенной заранее оговоренной периодичностью);
- по средствам коммуникации – устные и письменные услуги;
- по каналам коммуникации – локальные услуги (предоставляются в помещении научной библиотеки) и дистанционный режим.

Такой детальный подход к режимам оказания услуг имеет большое методическое значение для библиотек. Например, благодаря пониманию, что услуга, оказанная с помощью компьютера, не делает ее виртуальной, и тем более не делает виртуальным пользователя библиотеки. На самом деле пользователю библиотеки оказана обычная услуга, но в дистанционном режиме.

ГОСТ Р 7.0.104–2019 должен стать надёжным инструментом в оказании библиотечно-информационных услуг пользователям научных библиотек и органов научно-технической информации. Кроме того, на его основе возможна разработка локальных нормативных документов, определяющих содержание, формы реализации и режимы предоставления тех или иных услуг. Стандарт позволяет не только увидеть библиотечно-информационные услуги в перечне основных библиотечных процессов и операций, но и принимать во внимание сопутствующие факторы. Комплексный подход к разработке документа позволит повысить качество функционирования научных библиотек и органов НТИ. Несомненное достоинство стандарта заключается в привлечении теоретических и методических разработок из информатики, архивоведения и документоведения. Тем самым реально осуществляется системный подход в области национальной стандартизации, предусмотренный СИБИД. Стандарт не только содействует унификации библиотечных процессов и операций (как и их осмыслению), но и вносит значительный вклад в развитие отечественного библиотековедения. Хотелось бы надеяться, что данное направление стандартизации найдет дальнейшее развитие. Например, весьма актуальным представляется национальный стандарт, регулирующий условия для обслуживания специальных пользователей (наличие пандусов, освещение, специальная техника и т. д.). Не лишней была бы и разработка ГОСТа, определяющего обязательную номенклатуру услуг для библиотек каждого типа и вида, а также критериев их качественной оценки.

СПИСОК ЛИТЕРАТУРЫ

1. Об утверждении национального стандарта Российской Федерации: приказ от 18.01.19 №4-ст // Росстандарт: офиц. сайт. – URL: http://www/gost.ru/portal/gost/home/activity/ documents (дата обращения 25.06.19).
2. Об общероссийских классификаторах технико-экономической и социальной информации в социально-экономической области: постановление Правительства Российской Федерации от 10.11.2003 № 677 // Росстандарт: офиц. сайт. – URL: http://www/gost.ru/portal/gost/home/activity/classifandcatal (дата обращения 25.06.19).
3. ГОСТ 7.41-82 СИБИД. Единицы учета обслуживания читателей и абонентов библиотек и органов научно-технической информации // Электронный фонд правовой и информационно-технической документации: офиц. сайт. – URL:

http://www/docs.cntd.ru/document/1200004389 (дата обращения 25.06.19).
4. ГОСТ 7.0-99 СИБИД. Информационно-библиотечная деятельность, библиография. Термины и определения // Электронный фонд правовой и информационно-технической документации: офиц. сайт. – URL: http://www/docs.cntd.ru/document/1200004287 (дата обращения 25.06.19).
5. ГОСТ Р 7.0.20-2014 СИБИД. Библиотечная статистика. Показатели и единицы исчисления // Электронный фонд правовой и информационно-технической документации: офиц. сайт. – URL: http://www/docs.cntd.ru/document/1200113790 (дата обращения 25.06.19).
6. ГОСТ Р 7.0.103-2018 СИБИД. Библиотечно-информационное обслуживание. Термины и определения // Электронный фонд правовой и информационно-технической документации: офиц. сайт. – URL: http://www/docs.cntd.ru/document/1200161600 (дата обращения 25.06.19).
7. Десятый научно-практический семинар «Библиотечная поддержка исследований в сфере социальных и гуманитарных наук» // ИНИОН РАН: офиц. сайт: раздел «Семинары». – URL: http://www/inion.ru/scence/seminary/seminar-bibliotechnaia-podderzhka (дата обращения 25.06.19).

Материал поступил в редакцию 16.07.19

Сведения об авторах

ДЖИГО Александер Александрович – кандидат филологических наук, заведующий НИО библиотековедения, Институт научной информации по общественным наукам РАН, Москва
e-mail: adzhigo@hotmail.com

МАЙСТРОВИЧ Татьяна Викторовна – доктор педагогических наук, ведущий научный сотрудник, Институт научной информации по общественным наукам РАН
e-mail: t-maistr@yandex.ru

ВНИМАНИЮ ПОДПИСЧИКОВ!

С 2018 года возобновляется издание информационного бюллетеня «Иностранная печать об экономическом, научно-техническом и военном потенциале государств-участников СНГ и технических средствах его выявления» серии «Экономический и научно-технический потенциал» (56741) взамен информационного бюллетеня «Экономика и управление»

Периодичность выхода – 12 номеров в год. Объем 48 уч.-изд. л. в год.

В бюллетене освещаются материалы иностранной печати по широкому спектру вопросов, касающихся сфер экономического и научно-технического развития России и стран СНГ: общие вопросы, финансы, промышленность, рынки, сельское хозяйство, космос, транспорт и связь, природные ресурсы, трудовые ресурсы, внешние торгово-экономические и научные связи

Оформить подписку на информационный бюллетень, начиная с любого номера, можно в ВИНИТИ РАН по адресу: 125190, Россия, Москва, ул. Усиевича, 20,

Телефоны: (499) 151-78-61; (499) 155-42-85

Факс: (499) 943-00-60;

E-mail: contact@viniti.ru; sales@viniti.ru

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ РОССИЙСКОЙ АКАДЕМИИ НАУК

предлагает научным работникам, аспирантам и другим специалистам в области естественных, точных и технических наук, желающим быстро и эффективно опубликовать результаты своей научной и научно-производственной деятельности, использовать способ публикации своих работ через *систему депонирования.*

Депонирование (передача на хранение) – особый метод публикации научных работ (отдельных статей, обзоров, монографий, сборников научных трудов, материалов научных конференций, симпозиумов, съездов, семинаров), разрешенных в установленном порядке к открытому опубликованию.

Подготовка и передача на депонирование научных работ происходит в соответствии с «Инструкцией о порядке депонирования научных работ по естественным, техническим, социальным и гуманитарным наукам» (М., 2014).

Депонированные научные работы находятся на хранении в депозитарии ВИНИТИ РАН, копии работ предоставляются заинтересованным организациям и специалистам на бумажном и электронном носителях и являются официальной публикацией.

Информация о депонированных научных работах включается в информационные издания ВИНИТИ РАН: Реферативный журнал, Базу данных и Аннотированный библиографический указатель «Депонированные научные работы».

Направить научную работу на депонирование можно, обратившись в Группу депонирования ЦНИО ВИНИТИ РАН по адресу:

125190, Москва, ул. Усиевича, 20.
ВИНИТИ РАН, Группа депонирования ЦНИО
Тел.: 499-155-43-28, 499-155-43-76, 499-155-42-43, Факс: 499-943-00-60,
E-mail: cnio@viniti.ru, dep@viniti.ru

С инструкцией о порядке депонирования можно ознакомиться на сайте ВИНИТИ РАН: **http://www.viniti.ru**

ВНИМАНИЮ ЧИТАТЕЛЕЙ!

ВИНИТИ РАН, как единственный в России владелец лицензии Консорциума УДК, предлагает издания УДК полного четвертого издания на русском языке в печатном и электронном виде:

1. Таблицы УДК

УДК. Том I Общая методика применения УДК. Вспомогательные таблицы. Основные таблицы. Общий отдел. Алфавитно-предметный указатель к Общему отделу

УДК. Том II **1/3** Философия. Психология. Религия. Богословие. Общественные науки (только электронное издание)

УДК. Том III **5/54** Математика. Естественные науки (только электронное издание)

УДК. Том IV **55/59** Геологические и биологические науки (только электронное издание)

УДК. Том V **6/61** Медицинские науки (только электронное издание)

УДК. Том VI (часть 1) 6/621 Прикладные науки. Технология. Инженерное дело (только электронное издание)

УДК. Том VI (часть 2) 622/629 Техника. Инженерное дело (только электронное издание)

УДК. Алфавитно-предметный указатель к т. VI (1 и 2 части) (только электронное издание)

УДК. Том VII **63/65** Сельское хозяйство. Домоводство. Управление предприятием (только электронное издание)

УДК. Том VIII **66** Химическая технология. Химическая промышленность. Пищевая промышленность. Металлургия. Родственные отрасли (только электронное издание)

УДК. Том IX **67/69** Различные отрасли промышленности и ремесел. Строительство (только электронное издание)

УДК. Том X **7/9** Искусство. Спорт. Филология. География. История.

УДК. АПУ (с в о д н ы й) к полному 4-му изданию

УДК. Изменения и дополнения. Выпуск 2 (к т.т. 1–3) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 3 (к т.т. 1–6) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 4 (к т.т. 1–7) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 5 (к т.т. 1–10)

УДК. Изменения и дополнения. Выпуск 6 (к т.т. 1–10)

УДК. Изменения и дополнения. Выпуск 7 (к т.т. 1–10), 2017 г. (только электронное издание)

Для подписки необходимо направить заявку по адресу:

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

Телефоны: 499-155-42-85, 499-151-78-61

E-mail: feo@viniti.ru