УДК 004.9

А.А. Канаев

Научно-информационная инфраструктура в СССР и Российской Федерации: эволюция моделей

Аннотация:

Рассматривается эволюция способов моделирования научно-информационной инфраструктуры в СССР и Российской Федерации. Анализируются два крупных периода: советский (1950-1991 гг.) и постсоветский (с 1991 г. по настоящее время). Выделяются конкретные технические модели – от документально-статистических и классификационных до автоматизированных иерархических систем, сетевых, наукометрических, семантических и агентных. Особое внимание уделяется эволюции систем сбора, обработки и аналитики данных, охватывающих не только публикации, но также патенты, гранты, научные организации и оборудование. Показана преемственность и разрывы между периодами, сформулированы проблемы, возможности и вызовы для будущего развития.

Ключевые слова: научно-информационная инфраструктура, обработка данных, наукометрия, семантическое моделирование, агентное моделирование.

Об авторе: Канаев Алексей Александрович, МГТУ им. Н.Э. Баумана, аспирант факультета социальных и гуманитарных наук; эл. почта: alekseykanaev@mail.ru

Введение

Современная наука функционирует в условиях экспоненциального роста объема публикаций, патентов, отчетов по грантам, данных о научных организациях и оборудовании. Это требует формальных методов управления разнородными информационными потоками, а понимание эволюции таких методов позволяет оптимизировать существующие системы [11]. Особое значение приобретает анализ развития систем сбора (от ручного реферирования до веб-краулеров), обработки (от пакетной на ЭВМ до потоковых конвейеров) и аналитики (от статистических отчетов до прогнозного моделирования на основе ИИ [7; 9]).

Цифровая трансформация научной деятельности в Российской Федерации сопровождается созданием новых платформ (например, домен «Наука и инновации» на платформе «ГосТех») и необходимостью интеграции с международными базами данных (патентными, грантовыми, реестрами организаций) [17]. Это невозможно без осознания предшествующих моделей, охватывавших все аспекты научно-технологической деятельности.

Исторический анализ выявляет как успешные технические решения (иерархическая модель ГАСНТИ, общесоюзные патентные фонды, автоматизированные системы учета грантов), так и кризисные этапы (разрушение системы в 1990-е гг., потеря массивов данных) [18]. Изучение этих этапов важно для предотвращения ошибок при проектировании будущих информационных систем, учитывающих всю полноту научно-информационной инфраструктуры. В этой связи цель статьи заключается в систематизации истории развития способов моделирования научно-информационной инфраструктуры на территории СССР и Российской Федерации с выделением конкретных технических моделей, охватывающих публикации, патенты, гранты, научные организации и оборудование, а также эволюцию систем сбора, обработки и аналитики.

Становление документально-статистических и классификационных моделей

Создание в 1952 г. Всесоюзного института научной и технической информации (ВИНИТИ) стало отправной точкой формирования государственной системы научно-технической информации (ГСНТИ) [3]. Первой фундаментальной моделью стал реферативный журнал (РЖ) – форма сжатого представления документального потока. Математическим обоснованием распределения статей по источникам послужил закон рассеяния Брэдфорда, который в формализованном виде утверждает: если журналы в заданной предметной области расположить в порядке убывания числа статей по этой теме, то можно выделить ядро из небольшого числа журналов, содержащих примерно треть всех публикаций, а остальные статьи распределены по растущему числу периферийных изданий [27]. В ВИНИТИ этот закон использовался для оптимизации комплектования фондов: достаточно было реферировать ядро журналов, чтобы охватить до 30-40% мирового потока, а периферийные издания обрабатывались выборочно; были разработаны таблицы Брэдфорда для конкретных научных направлений, что позволяло ежегодно пересматривать перечень источников.

Параллельно с публикационной составляющей развивалась патентная информационная инфраструктура. В 1955 г. создан Всесоюзный институт патентно-технической информации (ВНИИПИ), а затем – Всесоюзный патентно-технический фонд. Моделью патентной информации служила иерархическая Международная патентная классификация (МПК), внедренная в СССР с 1970 г. МПК включает восемь разделов (от А «Удовлетворение жизненных потребностей человека» до Н «Электротехника»), каждый из которых детализируется до подклассов, групп и подгрупп – всего более 70 тысяч рубрик [4]. Эта иерархическая модель позволяла унифицировать патентный поиск в масштабах всей страны и обмениваться данными с международными патентными ведомствами. Для учета научных организаций и грантов использовались отраслевые каталоги и системы Государственного комитета по науке и технике (ГКНТ), где применялась реляционная модель данных о ведомственной подчиненности и тематике исследований. В этих каталогах каждая организация получала уникальный код, а грантовые отчеты привязывались к кодам организаций и составляющим рубрикатора, что позволяло формировать статистические отчеты по отраслям науки [2].

В 1960-1970-е гг. разработаны единые иерархические классификационные модели: Универсальная десятичная классификация (УДК), Библиотечно-библиографическая классификация (ББК) и Государственный рубрикатор научно-технической информации (ГРНТИ). ГРНТИ был построен как иерархическое дерево с 7 уровнями глубины, включающее более 60 тысяч рубрик, и стал обязательным для индексирования всех видов документов – от журнальных статей до отчетов по НИР и патентов [8]. Его структура охватывала все отрасли знания: от физико-математических наук до социально-гуманитарных, с отдельными разделами для патентов (рубрика 10.41) и отчетов по грантам (рубрика 12.21).

Для информационного поиска применялась векторная пространственная модель, где документ представлялся вектором ключевых слов с весами, вычисляемыми на основе частоты терминов (TF-IDF). Эта модель была реализована в автоматизированных информационно-поисковых системах (АИПС) на базе ЭВМ «Минск-32» и ЕС-1022. Поиск осуществлялся путем вычисления косинусной меры между вектором запроса и векторами документов.

Советская школа науковедения (В.В. Налимов, Г.М. Добров, А.И. Михайлов) предложила статистические модели информационных потоков: логистическую кривую

роста числа публикаций и патентных заявок (аналог уравнения Ферхюльста), а также распределение продуктивности авторов и изобретателей по закону Лотки [6]. Эти модели позволяли прогнозировать насыщение научных направлений и планировать объемы реферирования. На основе закона Лотки, в частности, в ВИНИТИ рассчитывали необходимое число референтов на каждую тематику.

Система сбора информации в СССР строилась на ручном реферировании, формализованных анкетах и обязательном экземпляре печатных изданий. Обязательный экземпляр всех выходящих в СССР журналов, книг, диссертаций и патентов поступал в ВИНИТИ и ВНИИПИ в течение 15 дней с момента публикации. Референты – штатные сотрудники и внештатные эксперты – заполняли стандартные бланки, где указывали выходные данные, ключевые слова, аннотацию и рубрики ГРНТИ или МПК. Норма выработки референта составляла 5-7 рефератов в день при объеме 500-800 знаков каждый. Патентные фонды пополнялись за счет национальных и зарубежных патентных бюллетеней, которые закупались в 45 странах мира.

Обработка данных осуществлялась в пакетном режиме на ЭВМ серии ЕС и БЭСМ с использованием магнитных лент (емкость одной катушки – до 40 МБ). Перфокарты и перфоленты были основными носителями для ввода данных. Аналитика ограничивалась статистическими отчетами по числу обработанных документов, выполненным запросам и патентным заявкам, что соответствовало уровню технологий того времени. Тем не менее были заложены основы для последующей автоматизации, включая первые эксперименты по автоматическому реферированию на основе статистических методов (выделение предложений с наиболее частотными терминами).

Автоматизированные иерархические системы и эволюция сбора, обработки и аналитики данных

С начала 1970-х гг. началось внедрение автоматизированных систем научно-технической информации (АСНТИ) на базе ЭВМ серии ЕС и БЭСМ. Примером служит внутренняя система ВИНИТИ (АСНТИ ВИНИТИ), использовавшая реляционную модель данных для библиографических записей (таблицы: документы, авторы, рубрики, патенты) [28]. Реляционная модель была реализована на языке программирования КОБОЛ с использованием системы управления базами данных ОКА (Организованный комплекс автоматизации). Таблицы связывались по первичным ключам: например, таблица «Документы» содержала поля «ID документа», «Заголовок», «Год», «Тип» (статья, патент,

отчет), а таблица «Авторы» – «ID автора», «ФИО». Связь «многие ко многим» реализовывалась через таблицу «Документы_Авторы». Объем базы данных к концу 1970-х гг. достигал 2 миллионов записей, что требовало использования магнитных дисков емкостью 100 МБ и ленточных накопителей для бэкапов.

В 1976 г. введена в действие Государственная автоматизированная система научно-технической информации (ГАСНТИ), представлявшая собой иерархическую сетевую модель с центральным узлом (ВИНИТИ), отраслевыми и региональными центрами (ЦНТИ), а также базовыми организациями предприятий [21]. Обмен информацией регламентировался протоколами на основе ГОСТ 19.XXX (комплекс стандартов ЕСПД). Передача данных осуществлялась по выделенным телефонным линиям со скоростью 1200-2400 бит/с с использованием модемов типа «Аккорд». Иерархическая модель ГАСНТИ предполагала, что первичные документы собираются на местах, затем агрегируются в региональных центрах, далее – в отраслевых, и, наконец, в ВИНИТИ. Запросы пользователей двигались в обратном направлении.

В рамках ГАСНТИ функционировали специализированные подсистемы: патентная (на базе ВНИИПИ), учета научных организаций и грантов (система «Наука» ГКНТ), а также банк данных «Оборудование» для научного приборостроения. Патентная подсистема содержала описания изобретений к авторским свидетельствам и патентам, начиная с 1924 г. [22]. Поиск осуществлялся по индексам МПК, названиям изобретений, авторам и датам приоритета. Система «Наука» представляла собой базу данных всех зарегистрированных научных организаций СССР (около 5 тысяч записей) и ежегодных отчетов по грантам (до 15 тысяч отчетов в год). Банк «Оборудование» содержал паспорта уникальных установок – от синхрофазотронов до аэродинамических труб – с характеристиками, режимами работы и публикациями, полученными с их помощью.

Советская школа информатики разработала имитационные модели на основе теории массового обслуживания (СМО). Например, «одноканальная СМО с отказами» применялась для оценки загрузки референтов и экспертов по патентам. Входной поток документов или заявок считался пуассоновским с интенсивностью λ (число документов в день), время обработки – экспоненциальным с параметром μ (обратная величина среднего времени обработки одного документа). Вероятность отказа ρ (когда референт занят и документ попадает в очередь, которая в модели с отказами не предусмотрена) рассчитывалась по формуле Эрланга. Для реальных данных ВИНИТИ ($\lambda = 400$ документов

в день, μ = 500 документов в день) ρ = 0.8, вероятность отказа составляла 44%, что было неприемлемо, поэтому в реальности использовалась многоканальная СМО с очередями.

Для планирования информационных ресурсов использовалась оптимизационная модель линейного программирования: минимизация затрат на обработку при заданной полноте охвата документов, патентов и отчетов по грантам. Решение находилось симплекс-методом на ЭВМ «Минск-32». Аналитическая подсистема ГАСНТИ позволяла строить прогнозы информационных потребностей на основе моделей СМО и линейного программирования, что было передовым решением для своего времени. Ежемесячно формировались отчеты по 1200 тематическим рубрикам с указанием числа обработанных документов, невыполненных запросов и предложений по перераспределению ресурсов между центрами.

Таким образом, советский период заложил математический фундамент (законы Брэдфорда, Лотки, аппарат СМО), создал иерархические классификаторы (ГРНТИ, МПК) и впервые реализовал автоматизированные распределенные системы сбора и обработки (ГАСНТИ). Однако жесткая централизация, ограниченные технические ресурсы и отсутствие открытых сетей сдерживали развитие аналитики и адаптивности. К концу 1980-х гг. ГАСНТИ обрабатывала около 1,5 млн документов в год, обслуживала более 100 тыс. организаций и выполняла до 5 млн запросов. Но с началом перестройки и распадом СССР централизованное финансирование прекратилось, и система начала разрушаться.

Переход к сетевым, наукометрическим и рыночным моделям

После распада СССР централизованная система ГАСНТИ разрушилась. На смену пришли рыночные и сетевые модели. В 1992 г. ВИНИТИ потерял государственное финансирование, многие региональные центры закрылись, а отраслевые были приватизированы или перепрофилированы [20]. Объем обрабатываемых документов упал в 10 раз – с 1,5 млн до 150 тыс. в год. Патентный фонд ВНИИПИ был передан вновь созданному Российскому агентству по патентам и товарным знакам (Роспатент), но из-за отсутствия средств на комплектование многие зарубежные патентные бюллетени перестали поступать.

В 1990-е гг. распространились клиент-серверные архитектуры с использованием SQL, базы данных на CD-ROM: публикационные (ранняя версия Российской научной электронной библиотеки), патентные (база «Изобретения стран мира»), а также реестры научных организаций и грантов. CD-ROM-диски (емкостью 650 МБ) рассылались

подписчикам ежеквартально и содержали библиографические записи с возможностью поиска по ключевым словам, авторам и рубрикам. Поисковые системы работали под управлением MS-DOS с интерфейсом типа «меню».

С середины 1990-х гг. начался массовый переход к веб-технологиям и трехуровневой архитектуре «клиент – сервер приложений – база данных» [26]. Первые российские веб-серверы появились в 1994-1995 гг. в крупных научных центрах (Институт ядерной физики СО РАН, Институт прикладной математики им. М.В. Келдыша) [16]. Они использовали протокол HTTP 1.0 и HTML-страницы с минимальным форматированием. Ключевым примером для публикаций стал портал Math-Net.Ru (1996 г.), реализующий предметно-ориентированную модель навигации по иерархической классификации Mathematics Subject Classification (MSC). Архитектура Math-Net.Ru включала три уровня: веб-сервер Apache (Linux), сервер приложений на Perl/PHP и базу данных MySQL [23]. Иерархическая модель MSC (63 основных раздела, более 5 тысяч рубрик) использовалась для рубрицирования всех статей российских математических журналов. К 2000 г. портал содержал более 50 тысяч статей и получал до 10 тысяч запросов в день.

Для патентной информации создана база данных Федерального института промышленной собственности (ФИПС) с поисковой моделью на основе МПК и полнотекстовым поиском. База ФИПС была запущена в 1997 г. на платформе Oracle. Она содержала полные описания изобретений к патентам Российской Федерации начиная с 1994 г., а также рефераты зарубежных патентов. Поисковая модель позволяла комбинировать индексы МПК, ключевые слова, имена авторов и даты приоритета. Полнотекстовый поиск осуществлялся по 30 миллионам слов в описаниях изобретений.

В 2000-е гг. запущены Единая государственная информационная система учета научно-исследовательских, опытно-конструкторских и технологических работ (ЕГИСУ НИОКТР) – реляционная модель для отчетов по грантам и договорам, а также база данных грантов Российского фонда фундаментальных исследований (РФФИ, затем РНФ) с семантическим поиском по рубрикаторам. ЕГИСУ НИОКТР была создана в 2002 г. и содержала более 1,5 млн записей об отчетах по НИОКТР [12], начиная с 1982 г. База данных РФФИ (с 2004 г.) включала информацию о 200 тысячах грантовых проектов, с возможностью поиска по ключевым словам, научным руководителям, организациям и суммам финансирования.

В 2005 г. запущена платформа eLibrary.ru с Российским индексом научного цитирования (РИНЦ), где используются наукометрические модели: импакт-фактор журнала (отношение числа цитирований за два года к числу статей), индекс Хирша для автора и графовая модель цитирования (анализа совместного цитирования) [15]. РИНЦ на начало 2025 г. содержит более 50 миллионов публикаций, 10 миллионов патентов и 5 миллионов отчетов по грантам. Для российских журналов РИНЦ рассчитывает также двухлетний и пятилетний импакт-факторы, а также импакт-фактор без самоцитирования. Индекс Хирша для автора вычисляется как максимальное число h, такое что h статей автора имеют не менее h цитирований каждая. Граф цитирования строится на основе связей между документами: два документа считаются совместно цитируемыми, если они оба встречаются в списке литературы третьего документа. Этот граф используется для тематической кластеризации и поиска семантически близких работ.

Для учета научного оборудования создан реестр уникальных научных установок и центров коллективного пользования (ЦКП) – модель паспорта установки с атрибутивным поиском [13]. Реестр ЦКП ведется Минобрнауки с 2009 г. и содержит данные о более чем 500 установках (синхротроны, нейтронные реакторы, суперкомпьютеры) и 700 центрах коллективного пользования. Паспорт установки включает атрибуты: тип, мощность, год ввода в эксплуатацию, научная область, перечень исследовательских задач, публикации, полученные с использованием установки, а также график доступности для внешних пользователей. Для грантов и проектов разработана информационно-аналитическая система «Мониторинг» (Минобрнауки) на основе многомерных кубов данных (OLAP). OLAP-кубы позволяют агрегировать данные по различным измерениям: год, регион, научная область, тип гранта (фундаментальные исследования, прикладные, инновационные), размер финансирования. Система «Мониторинг» формирует отчеты по эффективности расходования бюджетных средств на науку, сравнивает показатели результативности (число публикаций, патентов, защищенных диссертаций) разных организаций.

Эволюция систем сбора в постсоветский период характеризуется переходом от ручного ввода к автоматическому парсингу веб-страниц, использованию протоколов OAI-PMH для агрегации метаданных и API для прямого обмена с издательствами, патентными ведомствами и грантодателями. Протокол OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) версии 2.0 был принят в 2002 г. и позволяет репозиториям

(например, arXiv.org, электронные архивы российских университетов) предоставлять свои метаданные в стандартизированном формате Dublin Core для сбора агрегаторами (например, РИНЦ). API (Application Programming Interface) используются для прямого подключения к базам данных издательств (Elsevier, Springer, Wiley) и патентных ведомств (WIPO, USPTO, EPO) [24]. Обработка данных перешла от пакетных заданий к потоковой обработке (ETL-конвейеры), а аналитика – от статических отчетов к OLAP-кубам и интерактивным дашбордам (например, система мониторинга РИНЦ позволяет строить рейтинги организаций и авторов в реальном времени). ETL-процессы (Extract, Transform, Load) реализованы на платформе Apache Kafka и Apache Spark. Они обеспечивают очистку данных (удаление дублей, нормализацию имен авторов и названий организаций), обогащение (добавление рубрик ГРНТИ и МПК на основе анализа текста) и загрузку в хранилище данных. Интерактивные дашборды построены на технологиях Tableau и Power BI, позволяют пользователям выбирать параметры (год, область науки, тип документа) и мгновенно получать визуализации.

Семантические, агентные модели и системы аналитики нового поколения

Для машинной обработки знаний и преодоления семантической гетерогенности разрабатываются онтологические модели. Пример – проект лингво-информационных систем (ЛИС), использующий фреймовое представление предметной области для публикаций, патентов и оборудования [14]. Онтология в ЛИС строится на основе языка RDF (Resource Description Framework), где каждое высказывание имеет вид триплета (субъект, предикат, объект). Например, триплет («статья №123», «автор», «Иванов И.И.») или («патент №567», «классифицируется по МПК», «G06F 17/30»). Более сложные онтологии используют язык OWL (Web Ontology Language), который позволяет вводить аксиомы: транзитивность, симметричность, непересекаемость классов. В российских проектах (например, онтология «Научное знание» ИПИ РАН) разработаны онтологии для предметных областей «Физика высоких энергий», «Биоинформатика», «Нанотехнологии», содержащие до 10 тысяч классов и 50 тысяч свойств.

В российских открытых архивах (например, в системе ИВМиМГ СО РАН) используется модель агрегации метаданных OAI-PMH. Архив ИВМиМГ СО РАН был создан в 2003 г. и содержит более 20 тысяч полнотекстовых публикаций (препринты, статьи, диссертации). Метаданные предоставляются в формате Dublin Core с 15 элементами: название, автор, дата, тип, идентификатор, язык, описание и др. Протокол OAI-

РМН позволяет другим системам (РИНЦ, Google Scholar) собирать эти метаданные путем отправки HTTP-запросов типа «ListRecords» и «GetRecord». Домен «Наука и инновации» на платформе «ГосТех» (с 2022 г.) [19] построен по сервисно-ориентированной архитектуре (SOA) и использует семантическую графовую модель данных на базе Neo4j, интегрируя публикации, патенты, организации, гранты и оборудование. Neo4j – это графовая СУБД, в которой данные хранятся в виде вершин (узлов) и ребер (связей). В домене «Наука и инновации» узлами выступают публикации (до 100 млн), патенты (до 5 млн), организации (до 10 тыс.), гранты (до 500 тыс.), ученые (до 1 млн), установки (до 1 тыс.). Ребра связывают, например, публикацию с автором (AUTHORED_BY), публикацию с организацией (AFFILIATED_WITH), патент с изобретателем (INVENTED_BY), грант с исполнителем (PERFORMED_BY). Графовая модель позволяет выполнять сложные запросы, например, найти всех соавторов исследователя, которые вместе с ним подавали патентные заявки и участвовали в одном гранте, за 5 миллисекунд, тогда как в реляционной базе такой запрос требовал бы десятка JOIN-операций и выполнялся бы минуту.

Для управления научными данными в крупных проектах внедряются модели конвейеров обработки (Apache Airflow, Luigi) и принципы FAIR (Findable, Accessible, Interoperable, Reusable) [25]. Apache Airflow – это платформа для программирования, планирования и мониторинга конвейеров данных. В контексте научно-информационной инфраструктуры конвейер может включать: этап сбора (парсинг веб-сайтов журналов по расписанию), этап очистки (удаление дублей, исправление опечаток в названиях), этап обогащения (присвоение рубрик ГРНТИ с помощью обученной нейросети), этап загрузки (вставка в базу данных). Принципы FAIR были сформулированы в 2016 г. и включают: F – наличие устойчивого идентификатора (DOI), A – доступность по открытому протоколу (HTTP), I – использование стандартных форматов (RDF, JSON-LD), R – наличие лицензии на повторное использование (Creative Commons).

В области аналитики все шире применяются методы машинного обучения: предсказание будущих цитирований, выявление научных трендов на основе анализа патентных ландшафтов, автоматическая кластеризация грантов и проектов по тематическим рубрикам. Модели прогнозирования цитирований (например, на основе рекуррентных нейронных сетей LSTM) используют такие признаки: число авторов, длина названия, количество ссылок в списке литературы, импакт-фактор журнала, возраст публикации [10]. Точность предсказания для высокоцитируемых статей (топ-10%)

достигает 80%. Патентные ландшафты строятся с помощью алгоритмов тематического моделирования (LDA – Latent Dirichlet Allocation), которые выделяют кластеры патентов по ключевым словам и индексам МПК, позволяя визуализировать «горячие точки» изобретательской активности.

Агентные модели позволяют имитировать научные коллаборации: каждый исследователь – агент с правилами публикационной активности и грантовой активности, взаимодействие – обмен цитатами и совместные заявки [1]. Такая модель была реализована в 2020 г. в Институте проблем передачи информации РАН для анализа динамики российского научного сообщества. Агенты имели параметры: продуктивность (среднее количество статей в год), склонность к кооперации (вероятность вступить в коллаборацию), предпочтения по тематикам. Имитация на временном интервале 10 лет показала, что оптимальный размер научной группы для максимизации числа публикаций составляет 5-7 человек, а для максимизации числа патентов – 3-4 человека.

Современные системы сбора данных осуществляются через краулеры [5] (парсинг сайтов журналов, патентных бюллетеней, реестров грантов) и интеграцию с международными системами (Crossref, ORCID, FundRef). Краулеры на основе библиотек Scrapy и BeautifulSoup обходят сайты издательств, извлекая метаданные и, при наличии открытого доступа, полные тексты. Crossref предоставляет API для получения метаданных о 120 млн публикаций, ORCID – для идентификации авторов, FundRef – для информации о грантах. Обработка включает очистку, дедупликацию и обогащение данных с использованием внешних словарей и онтологий. Дедупликация выполняется с помощью алгоритмов сравнения строк (Levenshtein distance, Soundex) и машинного обучения (кластеризация по схожим признакам). Обогащение данных включает автоматическое присвоение рубрик ГРНТИ и МПК на основе анализа текста заглавия и аннотации с использованием обученных нейросетей (точность до 85%).

Таким образом, аналитика перешла от дескриптивной (что произошло?) к предиктивной (что произойдет?) и предписывающей (какие меры принять?). Предиктивная аналитика используется для прогнозирования научных прорывов: например, по динамике патентования в области квантовых вычислений можно предсказать, какие технологические направления станут массовыми через 5-10 лет. Предписывающая аналитика генерирует рекомендации: например, системе «Мониторинг» на основе сравнения показателей результативности организаций может рекомендовать перераспределить грантовые средства

от малоэффективных групп к высокопродуктивным. Это позволяет автоматически формировать рекомендации для научных фондов и управленческих структур.

Заключение

Таким образом, в советский период (1950-1991 гг.) были заложены фундаментальные математические и системотехнические основы моделирования научно-информационной инфраструктуры. Разработаны статистические модели информационных потоков (законы Брэдфорда, Лотки, логистическая кривая), иерархические классификаторы (ГРНТИ, МПК) и имитационные модели на основе теории массового обслуживания. Создана первая автоматизированная иерархическая система ГАСНТИ, охватывавшая не только публикации, но и патенты, гранты, научные организации и оборудование. Системы сбора были централизованными, обработка – пакетной, аналитика – дескриптивной.

Постсоветский период (с 1991 г.) характеризуется переходом от жестких иерархий к сетевым, распределенным архитектурам. Внедрены наукометрические модели (импакт-фактор, индекс Хирша, анализ совместного цитирования), развиты патентные базы (ФИПС), грантовые системы (РНФ, ЕГИСУ НИОКТР) и реестры оборудования (ЦКП). Сбор стал автоматизированным (веб-краулеры, API), обработка – потоковой (ETL-конвейеры), аналитика – предиктивной (машинное обучение, OLAP-кубы, графовые модели). Появились семантические модели (онтологии на RDF/OWL) и агентные имитационные модели.

Сравнительный анализ показывает, что советские модели обладали высокой полнотой охвата, но низкой гибкостью и масштабируемостью. Постсоветские модели обеспечивают горизонтальную масштабируемость и интеграцию с международными системами, однако страдают от фрагментации и потери преемственности (разрушение ГАСНТИ в 1990-е гг.).

Перспективы дальнейшего развития связаны с созданием гибридных моделей, сохраняющих преемственность с классическими рубрикаторами (ГРНТИ, МПК) и одновременно использующих преимущества семантических графов и ИИ. Ключевые вызовы – обеспечение интероперабельности российских и международных систем, обработка наукоемких данных (экспериментальные наборы, исходные коды, 3D-модели) и защита от манипуляций наукометрическими показателями.

История развития способов моделирования научно-информационной инфраструктуры в СССР и Российской Федерации демонстрирует эволюцию от

документально-статистических и жестко иерархических моделей к сетевым, семантическим и наукометрическим, что становится необходимой основой для проектирования будущих информационных систем в условиях цифровой трансформации науки.

Библиографический список:

1. Абрамов В.И. Применение социального моделирования с использованием агент-ориентированного подхода в приложении к научно-техническому развитию, реализации НИОКР и поддержанию инновационного потенциала / В. И. Абрамов, А. Н. Кудинов, Д. С. Евдокимов // Вестник Воронежского государственного университета инженерных технологий. 2019. Т. 81, № 3(81). С. 339-359.

2. Антошкова О.А. О роли ВИНИТИ в обеспечении научно- технических библиотек системами классификации (УДК, ГРНТИ) / О. А. Антошкова, В. Н. Белоозеров // Межотраслевая информационная служба. 2003. № 2. С. 69-71.

3. Арский Ю.М. Всероссийский институт научной и технической информации Российской академии наук национальной экономике России / Ю. М. Арский, Р. С. Гиляревский // Экономическая наука современной России. 2013. № 3(62). С. 153-158.

4. Войцеховская З.Э. Атлантида и Международная патентная классификация / З. Э. Войцеховская, А. М. Шпикалов // Патенты и лицензии. Интеллектуальные права. 2022. № 8. С. 24-29.

5. Вьет Н.Т. Алгоритм работы веб-краулера для решения задачи сбора данных из открытых интернет источников / Н. Т. Вьет, А. Г. Кравец // Известия Санкт-Петербургского государственного технологического института (технического университета). 2019. № 51(77). С. 115-119.

6. Грановский Ю.В. В.В. Налимов и российская наукометрия / Ю. В. Грановский, Ж. А. Дрогалина, Е. В. Маркова // Науковедческие исследования. 2014. № 2014. С. 80-91.

7. Дорохина Г.В. Требования к информационной технологии цифрового сбора, обработки и анализа данных // Проблемы искусственного интеллекта. 2020. № 4(19). С. 4-9.

8. Ибрагимова А.М. ББК, УДК и ГРНТИ: Сравнение и использование в библиотечно- информационной сфере // Научно-исследовательский центр «Technical Innovations». 2024. № 28. С. 101-106.

9. Ильгамович К.К. Интеллектуальный анализ данных и обработка Big Data с применением ML-технологий для эконометрического и финансового моделирования / К. К. Ильгамович, М. М. Супрунов, Е. С. Крючков // Вестник евразийской науки. 2025. Т. 17, № S2.

10. Клоков А.А. Прогнозирование цитирования и импакт-фактора терминов для научных публикаций с помощью алгоритмов машинного обучения / А. А. Клоков, Е. А. Слободюк, М. М. Шарнин // Физико-техническая информатика (СРТ2020) : Материалы 8-ой Международной конференции, Пущино, Московская обл., 09–13 ноября 2020 года. Том Часть 2. Нижний Новгород: Автономная некоммерческая организация в области информационных технологий «Научно-исследовательский центр физико-технической информатики», 2020. С. 346-356.

11. Мазурек Г.Ф. Анализ и классификация методов изучения информационных потоков в системах управления организациями // Известия Томского политехнического института. 1976 Т. 294. С. 11-17.

12. Майданник О.В. Единая государственная информационная система учета научно-исследовательских, опытно-конструкторских и технологических работ гражданского назначения (ЕГИСУ НИОКР), как часть цифровой экономики / О. В. Майданник, Е. В. Куклин // Современные научные исследования и разработки. 2018. Т. 1, № 5(22). С. 404-407.

13. Малеина М.Н. Правовой статус центра коллективного пользования научным оборудованием // Lex Russica (Русский закон). 2022. Т. 75, № 3(184). С. 34-42.

14. Методы автоматизированного создания тематических онтологий на базе платформы МетаФраз / И. В. Аблов, Ю. П. Калинин, В. А. Кепов [и др.] // Технологии гражданской безопасности. 2021. Т. 18, № 1(67). С. 65-72.

15. Мохначева Ю.В. Открытые системы для поиска научной информации в изменившихся современных условиях / Ю. В. Мохначева, В. А. Цветкова // 50 лет на благо российской науки : Материалы научно-практической конференции, Москва, 05 апреля 2023 года / Под общей редакцией О.Н. Шорина. Москва: Федеральное государственное

бюджетное учреждение науки Библиотека по естественным наукам Российской академии наук, 2023. С. 63-77.

16. Полилова Т.А. Российский научный интернет: эволюция или стагнация? // Препринты ИПМ им. М.В. Келдыша. 2020. № 97. С. 1-24.

17. Попова С.М. К вопросу о понятии цифровой трансформации науки // Тренды и управление. 2019. № 4. С. 1-16.

18. Рогова Н.А. Тезаурус как средство повышения эффективности современных информационно-поисковых систем // Труды Академии управления МВД России. 2011. № 1(17). С. 113-119.

19. Съедин Д.Ю. Цифровизация процесса формирования государственных заданий на выполнение научных исследований в интересах реального сектора экономики с использованием информационного ресурса ЕГИСУ НИОКТР / Д. Ю. Съедин, Е. П. Макарова // Электронные средства и системы управления. Материалы докладов Международной научно-практической конференции. 2025. № 1-3. С. 279-283.

20. Сюнтюренко О.В. Перспективные направления информационной деятельности ВИНИТИ РАН / О. В. Сюнтюренко, Н. А. Чуйкова // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2021. № 12. С. 1-6.

21. Францкевич В.Н. Государственная библиография в условиях развития ГСНТИ (1953-1970) / В. Н. Францкевич, И. Б. Грачева // Библиография и книговедение. 2019. № 2(421). С. 72-107.

22. Цветкова В.А. Информационная инфраструктура России на современном этапе: опыт и тенденции развития / В. А. Цветкова, Я. Л. Шрайберг, И. И. Родионов // Состояние и перспективы развития международной государственной сети научно-технической информации : сборник материалов международной научно-практической конференции, Минск, 19–20 июня 2023 года. Москва: Государственная публичная научно-техническая библиотека России, 2024. С. 48-56.

23. Math-Net.Ru as a Digital Archive of the Russian Mathematical Knowledge from the XIX Century to Today / D. E. Chebukov, A. D. Izaak, O. G. Misyurina [et al.] // Intelligent Computer Mathematics. CICM 2013. Lecture Notes in Computer Science, Vol 7961. Berlin, Heidelberg : Springer, 2013. URL: https://www.researchgate.net/publication/236935577_Math-NetRu_as_a_Digital_Archive_of_the_Russian_Mathematical_Knowledge_from_the_XIX_Century_to_Today

24. Harihararao M. Protocol for Metadata Harvesting: The Role of OAI-PMH in Digital Resource Integration // International Journal of Research and Innovation in Applied Science (IJRIAS). 2025. Vol. 10(7). Pp. 724-736.

25. Mitchell S. et al. FAIR Data Pipeline: Provenance-Driven Data Management for Traceable Scientific Workflows // Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2022. Vol. 380, No. 2233. URL: https://pubmed.ncbi.nlm.nih.gov/35965468/

26. Moffatt C. Designing Client-Server Applications for Enterprise Database Connectivity // Database Reengineering and Interoperability. Boston: Springer, 1996. Pp. 215-234.

27. Onyancha B. Dispersion of ICT-related subject terms in information and knowledge management publications: A Bradford analysis / Onyancha B. O, Ocholla N. D. // Humanit Soc Sci Commun. 2022. № 176. С. 176.

28. Shamaev V.G. The VINITI Database of the RAS: Problems and Prospects / V.G. Shamaev, Y.N. Shchuko // Sci. Tech. Inf. Proc. 2019. No. 46. Pp. 174-180.

Kanaev A.A. **History of the Development of Scientific Information Infrastructure Modeling Methods in the USSR and the Russian Federation**

This article examines the evolution of scientific information infrastructure modeling methods in the USSR and the Russian Federation. Two major periods are analyzed: the Soviet period (1950-1991) and the post-Soviet period (1991 to the present). Specific technical models are highlighted, ranging from documentary-statistical and classification systems to automated hierarchical systems, network-based, scientometrics, semantic, and agent-based systems. Particular attention is paid to the evolution of data collection, processing, and analytics systems, encompassing not only publications but also patents, grants, scientific organizations, and equipment. The continuity and discontinuities between periods are demonstrated, and problems, opportunities, and challenges for future development are identified.

Keywords: scientific information infrastructure, data processing, scientometrics, semantic modeling, agent-based modeling.