УДК 004.8

ОРКЕСТРАЦИЯ МЕТОДОВ АНАЛИЗА НАУЧНЫХ ДАННЫХ В ПРОЦЕССАХ РЕЦЕНЗИРОВАНИЯ

О. М. Атаева[1] [0000-0003-0367-5575], Н. П. Тучкова[2] [0000-0001-5357-9640]

[1, 2]Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва, Россия

[1]oataeva@frccsc.ru, [2]ntuchkova@frccsc.ru

Аннотация

Исследована проблема сочетания методов в задаче семантического анализа научных данных и публикаций при рецензировании. На разных этапах обработки данных в системе SciLibRu использованы различные методы, построена многоуровневая онтология, наполнен граф знаний, что приводит к формированию новой структуры данных, отличной от исходной. Каждый метод по отдельности приобретает свое назначение в такой системе, при этом в совокупности их сочетание приводит к возникновению новых свойств, которые стали предметом настоящих исследований. Приведен пример автоматического агента рецензирования с объяснимым результатом.

Ключевые слова: *оркестрация методов, семантический анализ, онтология предметной области, граф знаний, большие языковые модели, системы, категории, динамические структуры.*

ВВЕДЕНИЕ

Одним из парадоксов цифровизации стало то, что искусственному интеллекту (ИИ) «отдаются» творческие задачи, которые выполняют большие языковые модели (БЯМ), а человеку остаются рутинные задачи, такие как разметить текст, вычитать рукопись, исправить подписи к рисункам в статьях и т. д. Ученые продолжают активно учить БЯМ и восхищаются их способностям «рассуждать», но в то же время сетуют, что скоро труд исследователя обесценится и сведется к «правильной постановке вопроса» для БЯМ.

Проблема использования накопленной цифровой информации для сопровождения научных исследований по-прежнему остается актуальной. Еще один из парадоксов состоит в том, что методов обработки данных, которые как раз создавались для снятия рутинной нагрузки, огромное количество, но они не демонстрируют системного подхода.

Продукты ИИ, такие как система Prism OpenAI [1], российская разработка DoTrace [2] и др., ускоряют написание научных текстов, что вызывает вопрос о том, можно ли доверять этим результатам, считать ли их научными и как их рецензировать. Объем публикаций растет, инструменты множатся и при этом нарастает фрагментация: каждый метод решает свою задачу изолированно. Большое количество работы с данными не приводит к упорядочиванию творческого процесса исследователя и рецензента, а способствует росту хаоса в этой сфере (вспомним второй закон термодинамики и рост энтропии [3]).

Естественным становится желание вернуться к первоначальной задаче автоматизации работы с данными, использованию семантического анализа для извлечения знаний, применению всего комплекса методов для сопровождения научных исследований и рецензирования. Необходим переход к системной архитектуре, где граф знаний (ГЗ) может играть роль координационного ядра. Оркестрация методов служит такой цели и используется применительно к различным системам данных и управления [4], она связана с понятием самоорганизации и моделирования интеллектуальных агентов.

Мы будем обращаться к методам, предназначенным для представления знаний научной предметной области, для формального описания агентов и управления оркестрацией в интеллектуальной системе поддержки научных исследований. Для организации процессов предлагается использовать модель многоуровневой онтологии, которая организована по принципу разделения ответственности и включает пять взаимосвязанных уровней. Нижние уровни описывают объекты предметной области и информационные ресурсы. Средний уровень является ядром оркестрации, он представляет методы обработки как полноправные онтологические объекты с явно декларированными предусловиями и постусловиями. Верхние уровни описывают динамику поведения агентов с их возможностями и сам процесс оркестрации. Межуровневые отображения позволяют «рассуждателю» (системе логических рассуждений, reasoning

engine/system) выводить применимые методы автоматически, без жестко закодированной логики маршрутизации. Активизация различных методов при инициализации запросов с помощью БЯМ на естественном профессиональном языке задает динамику взаимодействия и самоорганизации модулей системы на основе формальных описаний, заложенных в архитектуре.

Под оркестрацией понимается как организация процессов в информационной системе на основе специальной структуры, где на разных уровнях управления данными применяются различные методы и/или их сочетание.

На примере задачи рецензирования научной публикации предложено описание цифрового агента формирования шаблона для эксперта.

1. БЛИЗКИЕ ПО ТЕМАТИКЕ ИССЛЕДОВАНИЯ

В современных информационных системах развиваются идеи интеграции на основе семантической оркестрации, эмерджентности, самоорганизации и динамических структур. Подходы самоорганизации порождают новые архитектуры, основанные на моделировании динамики интеллектуальных агентов. Ключевое направление составляют идея эволюции мультиагентных систем и переход от статических структур к динамическим, основанным на привлечении человека как эксперта в системах принятия решений.

В исследовании [5] предложена идея перехода от мультиагентной системы к интегрированной динамической адаптивной системе принятия решений, где интеллектуальные агенты функционируют совместно с экспертной оценкой человека. Авторы предлагают ввести социоэкономические метрики оценки эффективности системы оркестрации.

На фоне перечисленных тенденций изменяется сама организация взаимодействия программных модулей. Возникают системы, где управление ГЗ и онтологией инициализируется с помощью БЯМ. В работе [6] предложено несколько методологий построения ГЗ с применением БЯМ которые, в свою очередь, используют различные подходы на разных этапах процесса построения ГЗ. Дан обзор методов построения онтологий и ГЗ на основе динамических структур. Рассмотрены нисходящие и восходящие парадигмы моделирования онтологий, где БЯМ вносят изменения в формальные описания и коррекции при генерации он-

тологий. Нисходящая парадигма делает акцент на логике проектирования, используя БЯМ для преобразования входных данных на естественном языке в формальные онтологии таких стандартов, как OWL. Авторы работы [6] утверждают, что их подход значительно повышает согласованность и сложность взаимодействия в системе, используя метакогнитивные подсказки для обеспечения саморефлексии и структурной коррекции во время генерации онтологий.

Привлечение БЯМ к оркестрации методов не только на уровне взаимодействия моделей, но и на уровне построения онтологий приводит к тиражированию ошибок, которые выдают галлюцинации на реальных данных. Такие ошибки, которые получили название “bias” (предвзятость) [7], образуются при переводе и миграции понятий и терминов. Далее, при настройке на определенную предметную область происходят «смещение» или «предвзятость» понятий, которые невозможно устранить, поскольку они заложены в онтологии. Требуется переобучение БЯМ или изменение онтологии, т. е. применение дополнительных методов корректировки семантики информационных систем. В работе [7] и многочисленных источниках, на которые ссылаются авторы этого исследования, предложено на практике привлекать экспертов и/или дополнять знания тематическими данными, чтобы определять фрагменты “bias” и устранять их для повышения достоверности при поиске.

Вопросам семантики оркестрации посвящено исследование [8], где предложен инструментарий для формального описания условий корректности семантической оркестрации методов анализа в различных предметных областях. В [8] введено понятие логических условий (Validity Constraints, VC), которые должны выполняться в определенные моменты рабочего процесса анализа данных. Условия VC задают как инварианты корректности для промежуточных состояний конвейера обработки данных и как инструменты для обнаружения нарушений. В процессе оркестрации должно выполняться формальное задание ограничений и их проверка.

Заметим, что традиционная роль онтологии сводится к описанию предметной области. В условиях агентных интеллектуальных систем, где необходимо динамически выбирать варианты обработки запроса и компоновать разнородные методы семантического анализа от символьного SPARQL-поиска и векторного поиска по эмбеддингам до БЯМ-генерации и формальной верификации,

возникает потребность в дополнительном уровне – *онтологическом описании самих процессов и методов агента*.

Настоящая статья посвящена разработке многоуровневой онтологии, которая служит единой архитектурной основой для представления научных знаний предметной области и формального описания методов агента и управления их оркестрацией. Онтология в данном подходе первична по отношению к ГЗ. Граф представляет собой материализацию онтологии на конкретных данных предметной области, а оркестратор – функциональный компонент навигации по графу с помощью БЯМ и посредством онтологического рассуждения.

2. ОНТОЛОГИЧЕСКИЙ ПОДХОД К СЕМАНТИЧЕСКИМ НАУЧНЫМ БИБЛИОТЕКАМ

2.1 Источники, особенности и представление научных данных

Научные данные могут быть востребованы во многих задачах: это поиск, сравнение, структурирование, формализация и другие. Каждая задача требует своего метода, а данные могут быть различной природы: текст, формулы, изображения, графовые структуры и др. Оркестрация позволяет работать с этими и другими объектами информационной системы в единой структуре. Настоящее исследование посвящено оркестрации как сочетанию методов обработки, подобранных для конкретных задач и конкретных данных.

Чтобы понять масштаб поставленной задачи, необходимо оценить природу современных научных данных. Они больше не ограничиваются традиционными публикациями. Мы имеем классические статьи и монографии, энциклопедии и классификаторы, формальные библиотеки доказательств. Но все чаще рождаются материалы, созданные с участием ИИ: автоаннотации, синтетические обзоры, автоматически сгенерированные формализации.

Научные данные обладают рядом специфических свойств, которые делают их обработку особенно сложной. Перечислим эти свойства:

– узкоспециализированная терминология и сложные иерархии понятий. Одно и то же понятие может иметь разные интерпретации в различных контекстах;

– знание распределено между формальными и неформальными представлениями: доказательство может существовать одновременно в строгом формальном виде и в текстовом объяснении;

– научные области динамичны, появляются новые концепты, изменяются связи, возникают междисциплинарные направления;

– корпус знаний многоязычен и неоднороден по стилю и структуре. Эти особенности делают очевидным то, что простое сопоставление слов или векторная близость не обеспечивают полноценного понимания структуры знания;

– центральной структурой хранения и навигации по научным знаниям в современных интеллектуальных системах служит ГЗ, обеспечивающий структурированное представление сущностей предметной области и их отношений.

Для работы с классическими источниками была разработана семантическая библиотека LibMeta [9], где в основу модели данных была заложена онтология. Работа с искусственными источниками составляет новое направление, требующее установления достоверности и истинности утверждений. Опыт построения семантических библиотек, включая LibMeta [10] и ее новую версию SciLibRu [11], показывает, что ГЗ не может быть сконструирован в отрыве от онтологии. Именно онтология предметной области определяет структуру данных, типологию понятий и семантически значимые связи, на основе которых граф строится и пополняется. Принцип «сначала онтология и тезаурус предметной области, затем граф знаний» является методологической основой такого подхода. Данные и технология их интеграции описаны в публикациях авторов [12, 13].

В настоящей работе представлена архитектура многоуровневой онтологии представления научных знаний, описания методов и их оркестрации. Оркестратор представляет собой функциональный компонент навигации по графу посредством онтологического рассуждения в многоуровневой онтологии [14].

Знания библиотек LibMeta и SciLibRu относятся к разделам математики. В работе авторов [15] предложена онтология SciLib на языке OWL/DL. Проведена материализация ГЗ на данных MathLib и Lean 4 [16]: построена таксономия доменов, выполнено сопоставление объектов с классами онтологии и сформированы мультимодальные RDF-представления для исторических и искусственных данных.

2.2. Многоуровневая онтология O: формальное описание и принципы построения

Существующие онтологические подходы к организации семантических библиотек описывают структуру научного знания: концепты предметной области, тезаурусы, термины, информационные ресурсы и их взаимосвязи [17–20]. Этого достаточно для хранения и навигации, но недостаточно для интеллектуального ассистента (программного цифрового агента), который должен выбирать и применять методы обработки данных, адаптируясь к состоянию задачи. Агент, располагающий лишь онтологией предметной области, «знает», о чем идет речь, но «не знает», как действовать и с помощью какого метода отвечать на конкретный информационный запрос [21]. Отсюда вытекает центральное требование к структуре информационной системы: *онтология системы должна описывать не только объекты мира, но и процессы работы агента, такие как методы обработки, их предусловия и постусловия, возможности агентов и правила их взаимодействия*. При таком описании оркестрация методов может быть выведена на основе правил, а не закодирована жестко в программном коде.

Предлагаемая многоуровневая онтология O строится по принципу разделения ответственности: каждый уровень описывает строго определенный аспект системы, а межуровневые отображения задают, как знания одного уровня управляют поведением на следующем [5]:

$$O = \langle L_1, \dots L_5 \rangle.$$

Каждый уровень L_k $(k = 1, \dots, 5)$ реализован как OWL 2 DL-онтология со своими классами, объектными свойствами и аксиомами. Межуровневые связи задаются отображениями $\phi_k: L_k \to L_{k+1}, (k = 1, \dots, 4)$, реализованными через именованные объектные свойства OWL. Это позволяет при выполнении рассуждений реализовать вывод через границы уровней, начав с объекта предметной области (L_1), пройти через представление (L_2), метод (L_3), агента (L_4) и достичь решения об оркестрации (L_{51}) без жестко закодированной логики маршрутизации. Нижние уровни L_1 и L_{12} описывают устойчивые (в смысле описания предметной области) объекты: концепты предметной области и информационные

ресурсы. Уровень L_3 является ядром оркестрации, он описывает методы как онтологические объекты первого уровня. Уровни L_4 и L_5 описывают динамику для агентов с их возможностями и сам процесс координации [15].

Из сказанного следуют четыре принципа, положенные в основу проектирования предлагаемой онтологии.

Принцип* 1. Первичность онтологии.* Онтология первична по отношению к ГЗ. Граф есть материализация онтологии на конкретных данных предметной области, а оркестратор – функциональный компонент навигации по графу посредством онтологического рассуждения [14]. Принцип «сначала онтология и тезаурус предметной области, затем граф знаний» является методологической основой данного подхода.

Принцип* 2. Разделение ответственности.* Каждый уровень онтологии описывает строго определенный аспект системы: предметную область, информационные ресурсы, методы обработки, агентные возможности, процесс оркестрации. Такое разделение восходит к архитектурному паттерну *Separation of Concerns* и обеспечивает модульность: изменение одного уровня не требует переработки остальных.

Принцип* 3. Выводимость оркестрации.* Межуровневые отображения связывают уровни таким образом, что система логических рассуждений может автоматически вывести, какой метод применим к данному объекту в данном состоянии, без жестко закодированной логики маршрутизации. Это существенно отличает предлагаемый подход от традиционных API-каталогов и статических конвейеров.

Принцип* 4. Единство представления.* Все уровни реализуются в рамках одного формализма OWL 2 DL. Это позволяет использовать стандартные OWL-рассуждатели (OWL Reasoner) для вывода через границы уровней и обеспечивает совместимость с существующими инструментами семантического веба (SPARQL, SHACL, SWRL).

На рис. 1 представлена иерархическая схема многоуровневой онтологии, где на каждом уровне обозначены содержание и связи в виде отображений $\phi_k: L_k \to L_{k+1} (k = 1, \dots, 4)$.

Рис. 1. Схема многоуровневой онтологии.

2.3. Уровни онтологии

2.3.1. *Первый уровень L_1: онтология предметной области.*

Первый уровень описывает структуру знания в конкретной научной области в рамках библиотеки SciLibRu, формализуя концепты, их иерархии, тематические классификаторы, формулы и тезаурусные отношения:

$$L_1 = \langle C_D, R_D, A_D, TH \rangle.$$

Здесь C_D – множество классов предметной области (*Concept, Formula, Domain, MathStatement*); R_D – объектные и аннотационные свойства, включая таксономические и содержательные отношения; A_D – аксиомы OWL 2 DL, задающие ограничения кардинальности и цепочки свойств.

Тезаурус $TH = \langle T, R_{TH} \rangle$ является неотъемлемой частью уровня: T – множество терминов, R_{TH} – иерархические и горизонтальные отношения между ними [22].

Связи между тезаурусными терминами и информационными объектами задаются семью семантически значимыми отношениями P_1–P_7. Отношения

$P_1(t, io)$ и $P_2(io, t)$ устанавливают двунаправленные связи между терминами тезауруса и информационными объектами. Отношение $P_3(r, s)$ связывает тип информационного ресурса с классом исходных объектов, $P_4(a, sa)$ – атрибут ресурса со свойством исходного класса. Отношения P_5, P_6, P_7 описывают соответственно принадлежность объекта классу источника данных, связь семантической метки с объектом и обратную связь объекта с меткой. OWL-рассуждатель расширяет ГЗ производными триплетами на основе аксиом, что и обусловило 310-кратный рост числа триплет при материализации Mathlib в SciLibRu [15, 16].

2.3.2. ***Второй уровень*** L_2***: онтология источников и ресурсов.***

Второй уровень описывает информационные ресурсы независимо от их предметного содержания. Назначение этого уровня – зафиксировать, в какой форме и из каких источников данных поступает информация в систему:

$$L_2 = \langle C_R, R_R, A_R \rangle.$$

Классы C_R включают *Publication, Monograph, Encyclopedia, Dataset, KnowledgeGraph, Corpus, Formula*. Свойства R_R описывают характеристики ресурсов: формат, язык, источник происхождения, лицензию, версии. Аксиомы A_R задают ограничения целостности, в частности, что каждый ресурс имеет не более одного канонического URI (Uniform Resource Identifier, унифицированный идентификатор ресурса).

Ключевым структурным элементом L_2 является многомодальная модель представления данных в онтологии SciLibRu. Уровень интерпретации описывает абстрактный смысл объекта, например понятие «теорема Пифагора» как математической «истины» независимо от ее конкретного выражения в виде текста или формулы. Уровень представления фиксирует конкретную форму: формальный код на Lean 4, текст на естественном языке, запись в LaTeX-нотации или визуализацию формулы. Уровень ресурса соответствует материальному носителю, то есть файлу, записи в базе данных или исполняемому коду. Все три подуровня связаны через общий URI. Такое разделение реализует принцип инвариантности: добавление новой модальности представления не изменяет интерпретационный уровень и не требует переработки L_1. Отображение $\phi_1: L_1 \to L_2$ реализуется свойством *hasRepresentation* [15].

2.3.3. Третий *уровень* L_3*: онтология методов обработки.*

Третий уровень является ядром оркестрации методов. Каждый метод семантического анализа и обработки данных представлен как самостоятельный онтологический объект с явно описанными условиями применимости. Это принципиально отличает L_3 от традиционных API-каталогов (Application Programming Interface): метод в L_3 – это не просто вызываемая процедура, а концепт с семантикой, над которым работает рассуждатель:

$$L_3 = \langle C_M, R_M, \mathrm{pre}(m_i), \mathrm{post}(m_i) \rangle.$$

Иерархия классов C_M строится от абстрактного *SemanticMethod* к конкретным подклассам: *QueryMethod* (*SPARQL, Cypher, GraphQL*), *EmbeddingMethod, GraphTraversalMethod, VerificationMethod, NLGenerationMethod, IndexingMethod*.

Свойства R_M включают *hasPrecondition, hasPostcondition, hasInputType, hasOutputType, hasComputationalCost, isComposableWith* и *conflicts*.

Каждый конкретный метод m_i формализуется как именованный индивид OWL. Его предусловие $\mathrm{pre}(m_i)$ – это конъюнкция утверждений об элементах L_1 и L_2, которые должны быть выполнены до вызова; его постусловие $\mathrm{post}(m_i)$ – это гарантированные изменения в состоянии системы после успешного исполнения. Отображение $\phi_2 \colon L_2 \to L_3$ реализуется свойством *isProcessedBy*.

2.3.4. Четвертый *уровень* L_4*: онтология агентных процессов.*

Уровень L_4 отвечает на вопрос: *кто* выполняет методы из L_3, *какими* возможностями располагает, *в каком* состоянии находится и *какую цель* преследует. Если L_3 описывает методы как абстрактные онтологические объекты с предусловиями и постусловиями, то L_4 связывает эти методы с конкретными исполнителями – программными агентами, каждый из которых обладает ограниченным набором инструментов и действует в рамках определенной политики (в рамках онтологии уровня). Разделение L_3 и L_4 реализует *Принцип* 2: добавление нового агента не требует изменения описания методов, а добавление нового метода – перепроектирования агентов, достаточно обновить множество инструментов агента.

Четвертый уровень L_4 описывает агентов как структурированные онтологические объекты, отвечая на вопрос: *кто* выполняет методы из L_3, *какими* возможностями обладает, в каком *состоянии* находится и какую *цель* преследует:

$$L_4 = \langle C_A, R_A, A_A \rangle,$$

где классы C_A включают *Agent, Capability, Tool, AgentState, Intent, Plan*.

Агент α формально задается кортежем

$$\alpha = \langle \mathrm{cap}(\alpha), \mathrm{tools}(\alpha), s_0, \pi_\alpha \rangle,$$

где $\mathrm{cap}(\alpha) \subseteq C_A$ – множество его возможностей (функций), $\mathrm{tools}(\alpha) \subseteq C_M$ – набор методов из L_3, s_0 – начальное состояние, $\pi_\alpha : S \to C_M$ – политика (правила) выбора метода.

Особую роль играют классы *Intent* и *Plan*:

– *Intent* описывает цель агента в терминах предметной области L_1 и связан с методами через свойство *requiresMethod*;

– *Plan* представляет собой частично упорядоченное множество функций с ограничениями порядка *hasPrecedence*, задающее декларативную последовательность достижения сложной цели [15, 16, 22].

2.3.5. Пятый *уровень* L_5*: онтология оркестрации.*

Уровень L_5 является метауровнем всей системы. Если L_1–L_4 описывают, *что* есть в мире (в рамках знаний системы), *в какой форме*, *как* обрабатывается и *кем* выполняется, то L_5 описывает сам *процесс координации*: как оркестратор принимает решения, как фиксирует их обоснование, как реагирует на неудачи. Ключевая функция L_5 – это обеспечение полной трассируемости цепочки рассуждений (explainability): каждое решение оркестратора привязано к конкретному SWRL-правилу, каждому состоянию до и после применения метода, каждому объекту предметной области. Без этого свойства нейросимволический агент не имеет преимуществ перед чисто нейронной системой в части объяснимости [22]:

$$L_5 = \langle C_O, R_O, A_O \rangle,$$

где классы C_O включают *OrchestratorAgent, TaskState, MethodSelection, ExecutionTrace, FallbackStrategy*. Ключевым классом является *ExecutionTrace*: каждый его

экземпляр фиксирует выбранный метод, состояние до и после его применения, а также правило онтологического вывода, обосновавшее выбор, что обеспечивает полную трассируемость цепочки рассуждений – требование *explainability*, реализованное в SciLibRu.

Пять уровней образуют связную цепочку оркестрации.

3. ОРКЕСТРАЦИЯ МЕТОДОВ

Методы, получаемые при структурировании данных, соответствуют уровням онтологического проектирования.

– На *формально-синтаксическом* уровне фиксируются объекты, заданные в формальных языках и системах доказательства: определения, леммы, теоремы, правила вывода, тактики, типы, а также их зависимости. В случае математики это может быть структура библиотеки, подобной MathLib, в которой явно представлены и формальные формулы, и дерево их использования. В других дисциплинах аналогичную роль играют формальные спецификации моделей, алгоритмов или протоколов.

– *Теоретико-структурный* уровень поднимается над конкретными формулами и описывает теории, разделы, математические и научные структуры, а также отношения между ними: обобщение и специализацию, эквивалентность формулировок, сведение одной теории к другой, импорты понятий и результатов. Здесь отдельные утверждения группируются в более крупные смысловые блоки такие, например, как теория меры, теория вероятностей, функциональный анализ, а в экспериментальных науках это определенные экспериментальные парадигмы и классы моделей.

– *Методологический* уровень фокусируется на процессуальной стороне научного знания. Он описывает методы доказательства, типичные шаблоны рассуждений, экспериментальные дизайны, протоколы сбора и анализа данных, стратегии выбора моделей и критериев. Важной частью этого уровня является описание «ролей» отдельных лемм, фактов и шагов в доказательствах или исследованиях: какие из них являются ключевыми, какие выполняют техническую вспомогательную функцию, какие обеспечивают переход между теоретическими и эмпирическими слоями.

– *Объяснительный, или дидактический*, уровень связывает всю эту формальную и структурную сложность с человеческим пониманием. На нем онтология фиксирует связи с энциклопедическими статьями, учебниками, обзорными материалами и примерами, а также задает различные стили и глубины объяснения одного и того же результата для разных аудиторий от начального уровня до экспертов. Это позволяет использовать один и тот же семантический каркас как в исследовательских, так и в образовательных сценариях.

3.1. LibMeta: оркестрация методов семантической библиотеки

Онтология LibMeta организует знания вокруг трех групп концептов: концептов, описывающих содержание предметной области и образующих тезаурус; концептов, описывающих тематические коллекции, и концептов, поддерживающих интеграцию данных из внешних источников. Тезаурус построен в соответствии со стандартом ISO 25964, включает иерархические отношения BT/NT и горизонтальное отношение RT. Граф KG MathSemanticLib строится итерационно, начиная с нулевой версии на основе Математической энциклопедии И. М. Виноградова [23]. Формулы в LibMeta являются полноправными семантическими объектами с собственными ребрами ГЗ к концептам и публикациям [9, 10, 22].

Пример навигации по предметной области обыкновенных дифференциальных уравнений иллюстрирует пайплайн с применением различных методов на рис. 2.

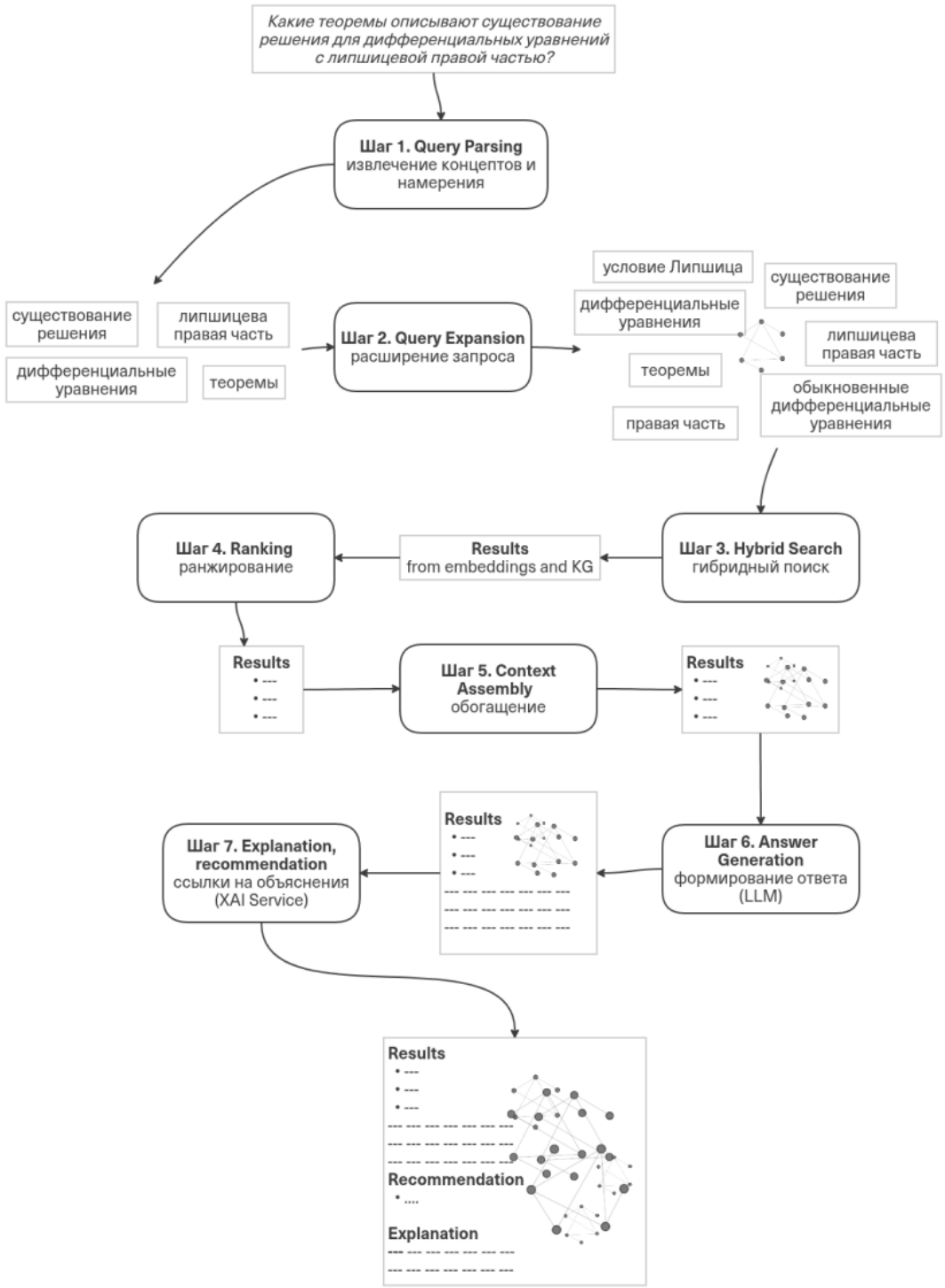

Рис. 2. Схема оркестрации методов на примере поиска в LibMeta.

При запросе «Уравнение Бернулли» оркестратор идентифицирует концепт *BernoulliODE*. Через тезаурусные связи извлекаются родственные концепты: уравнения Риккати и Якоби, коды классификации MSC, монографии и учебники

(см. рис. 2). Без онтологически управляемого контекста ведущие языковые модели (ChatGPT 4, YandexGPT) возвращают лишь общую информацию без формул и ссылок; ответ LibMeta трассируется (прослеживается в системе пошагово) до конкретных ребер графа и первоисточников [22].

Оркестрация реализуется как многошаговый пайплайн обработки запроса. На шаге разбора запроса (*query parsing*) определяются тип запроса и связанный концепт. На шаге расширения (*query expansion*) оркестратор применяет онтологические операции: подъем по ребру BT к более общему концепту или сужение по ребру NT, используя структуру тезауруса как механизм управления областью поиска, а не как классификатор пользователя. На шаге гибридного поиска (*hybrid search*) расширенный запрос обрабатывается параллельно: SPARQL-запросом к RDF-графу и векторным поиском по эмбеддингам. На шаге сборки контекста (*context assembly*) объединенный результат передается языковой модели. На шаге интеграции с XAI каждое утверждение связывается с конкретной публикацией или статьей энциклопедии через ребра графа. Это требование принципиально для математических предметных областей: недостаточно получить правильный ответ, необходимо верифицировать его через первоисточник.

3.2. SciLibRu: развитие подхода для формальной математики

Конкретным применением и развитием оркестрации методов является материализация библиотеки Lean 4 Mathlib в RDF-граф путем интеграции данных, утверждений и источников. Пайплайн материализации включает компиляцию Lean 4, извлечение метаданных парсером JIXIA, автоматическую аннотацию 660 тематических подклассов класса Domain, отображение в модель SciLib и материализацию в GraphDB. Результирующий граф содержит 66 млн RDF-троек с 6.3 млн уникальных субъектов, позволяет реализовать 310-кратный прирост относительно исходных объявлений Mathlib, достигнутый через онтологическую типизацию и материализацию выводов рассуждателя.

В SciLibRu оркестрируемые методы – это восемь режимов доставки контекста, использующих одну базовую модель DeepSeek-Prover-V2-7B и различающихся только способом формирования контекста. Первый базовый режим передает модели исходную формулировку без подсказок из графа. Второй режим

дополняет контекст леммами из векторного поиска. Следующие режимы используют граф зависимостей Mathlib с различными стратегиями обхода. Последние два гибридных режима объединяют структурный обход графа с векторным поиском и достигают наилучших результатов. На 109 трудных задачах miniF2F-Test лучший гибридный режим почти утраивает базовый показатель и вероятность того, что модель сгенерирует правильный код, с первой попытки растет с 3.56% до 10.42% [15, 24].

Принципиально важный результат заключается в том, что детерминированные символьные правила превосходят нейронные методы выбора точек входа в граф. Девять регулярных выражений, сопоставляющих структурные признаки Lean 4 с фиксированными точками входа в граф Mathlib, выполняются за 12 с и не требуют дополнительных вызовов БЯМ. Нейронный аналог требует около 30 вызовов и 134 с при более низкой точности. Этот результат имеет прямое значение для архитектуры оркестратора: для навигации по структурированному ГЗ декларативные символьные правила работают надежнее, быстрее и дешевле, чем нейронная генерация.

3.4. Общие закономерности LibMeta и SciLibRu

Анализ LibMeta и SciLibRu в их преемственности позволил выделить три устойчивые закономерности, согласованные между собой.

Первая закономерность: онтология наиболее эффективна при дефиците параметрических знаний. В LibMeta специализированные российские математические источники недоступны открытым БЯМ. В SciLibRu на легких задачах граф-расширение не дает значимого прироста; на трудных задачах, где модели не хватает «словаря» тактик, граф почти утраивает успешность. Внешнее знание ценно именно там, где внутреннего (в рамках данных и источников библиотеки) не хватает.

Вторая закономерность: символьные правила эффективнее нейронной генерации для навигации по структурированному графу. Тезаурусные операции BT/NT в LibMeta и regex-паттерны в SciLibRu являются детерминированными символьными правилами, в обоих случаях они работают быстрее и точнее. Когда пространство поиска уже структурировано онтологией, дополнительное нейронное рассуждение для навигации по нему избыточно.

Третья закономерность: трассируемость является необходимым, а не опциональным свойством. В LibMeta каждое утверждение привязано к публикациям и энциклопедическим статьям; в SciLibRu – к ребрам *usesInType/usesInValue* графа Mathlib. Без этого свойства нейросимволический агент не имеет преимуществ перед чисто нейронной системой в части объяснимости.

4. ПРИМЕР: ЦИФРОВОЙ АГЕНТ ПОСТРОЕНИЯ ШАБЛОНА ДЛЯ РЕЦЕНЗЕНТА

Рецензирование научных работ составляет одну из наиболее востребованных сфер деятельности эксперта. Это часть подготовки публикаций и докладов, которая требует времени, но в ограниченном временном промежутке, в соответствии с многочисленными дедлайнами. В классической работе [25], где перечислены 73 этапа работы с научным журналом, автор отмечает, что есть проблема «отклонение статьи». Это одно из противоречий издательского дела, поскольку статьи нужны, но нельзя пропустить ошибочные результаты и не учесть остальные особенности научных публикаций, такие как новизна, актуальность и др. Роль рецензента научных публикаций остается существенной. Тем не менее, как правило, есть определенные требования к составлению отзывов (рецензий), которые можно формализовать в виде шаблонов и далее предоставить эксперту предварительно подготовленный шаблон для анализа и работы.

Пример использования агентного подхода [26, 27] к решению проблемы рецензирования предлагается в нашем исследовании на основе оркестрации методов анализа данных (см. рис. 3). Связи ($ф_i$, $ф^{-1}_i$) отвечают за связи между уровнями онтологии при кординации действий агента. Например, пара связей (*orchestrates, getOrchestrated*) описывает выбор серии действий и проверку их результата, (*canExecute, getExecuted*) – проверку отдельных действий и их предварительных условий и проверку постусловий для них, (*isProcessedBy, getProcessedBy*) – применение отдельных методов, (*hasRepresentation, getRepresentation*) – представление извлеченных результатов с помощью этих методов.

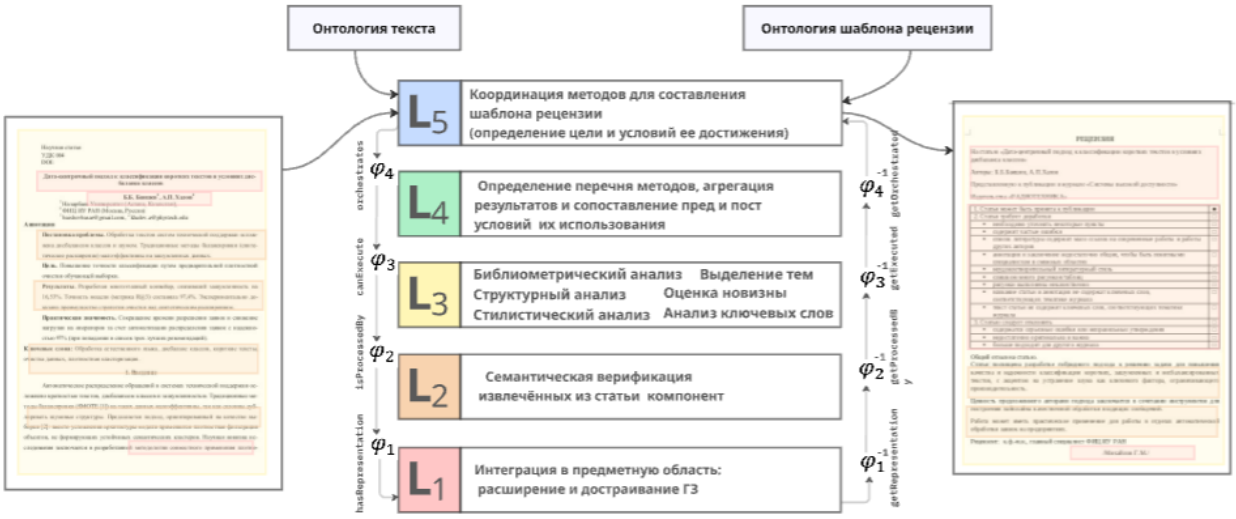

Рис. 3. Схема связей цифрового агента.

На рис. 3 показано, как шаблон рецензии ReviewTemplate заполняется поэтапно после обработки исходного текста научной статьи из раздела математических предметных областей на разных уровнях онтологии $O = \langle L_1, L_2, L_3, L_4, L_5 \rangle$:

L_1 – интеграция текста научной статьи в предметную область (используется описание предметной области, достраивается ГЗ);

L_2 – семантическая верификация (выявление метаданных, структуры текста и концептов предметной области);

L_3 – библиографическая верификация (применение методов обработки теста для структурного анализа, верификации формул, анализа ключевых слов, стилистического анализа, анализа библиографии);

L_4 – определение перечня методов (активация агента рецензирования);

L_5 – координация методов для составления шаблонов (организация конвейера для заполнения шаблона).

Результатом реализации схемы рис. 3, становится заполненная форма шаблона, которая предоставляется рецензенту для принятия решения о публикации.

ЗАКЛЮЧЕНИЕ

Идея создания многоуровневой онтологии научных данных и ее материализации в виде семантического научного графа задает основу для построения

научно-информационных систем нового поколения. В таких системах данные и знания перестают быть разрозненным набором файлов, записей и кодов; они становятся встроенными в согласованное семантическое пространство, поддерживаемое онтологией и графом знаний. Вокруг этого пространства разворачивается инфраструктура интеллектуальных сервисов, позволяющая не только находить и интегрировать информацию, но и формально рассуждать, объяснять решения ИИ-моделей и поддерживать обучение и исследовательскую работу. В результате формируется новый тип научной динамической инфраструктуры, в которой объяснимый ИИ выступает не внешней надстройкой, а частью системы, опирающейся на богатую, многоуровневую модель научного знания и ее инженерную реализацию.

Линия исследований LibMeta – SciLibRu формирует конкретную траекторию направления, которое можно обозначить как *онтологически управляемое нейросимволическое ассистирование научных исследований*. Их конвергенция к общим архитектурным принципам, трассируемости каждого решения к структуре онтологии, приоритету символьных правил для навигации по структурированному пространству, инвариантности онтологии при добавлении новых данных дает основание полагать, что эти принципы отражают фундаментальные свойства надежных нейросимволических агентов в научной среде.

СПИСОК ЛИТЕРАТУРЫ

1. PRISM. https://openai.com/ru-RU/prism/ (дата обращения 14.04.2026).

2. DoTrace. https://www.domate.ru/dotrace_platform (дата обращения 14.04.2026).

3. *Кубо Р.* Термодинамика. М.: Мир, 1970, 304 с.

4. A Unified Framework for Self-Organizing Intelligence: A Synthesis of Computational Autopoiesis, Category Theory, and Iterative Concept-Abstraction Cycles**.** Academia.edu. SSRN. 2025. https://www.academia.edu/143199301/ (дата обращения 14.04.2026).

5. *Tallam K.* From Autonomous Agents to Integrated Systems, A New Paradigm: Orchestrated Distributed Intelligence // arXiv:2503.13754. 2025. https://doi.org/10.48550/arXiv.2503.13754

6. *Bian H.* LLM-empowered knowledge graph construction: A survey // arXiv:2510.20345. 2025. https://doi.org/10.48550/arXiv.2510.20345

7. *Zabihi P., Nawara D., Ibrahim A., Kashef R.* Analyzing Bias in LLM-Augmented Knowledge Graph Systems: Taxonomy, Interaction Mechanisms, and Evaluation // Applied Sciences. 2026. Vol. 16, No. 7. Art. 3410. https://doi.org/10.3390/app16073410

8. *Schintke F. et al.* Validity constraints for data analysis workflows // Future Generation Computer Systems. 2024. Vol. 157. P. 82–97. https://doi.org/10.1016/j.future.2024.03.037

9. *Ataeva O.M., Serebraykov V.A., Tuchkova N.P.* Approaches to the organization of mathematical knowledge when forming subject thesauruses of various mathematics domains // CEUR Workshop Proc. 2018. Vol. 2260. P. 42–54. https://doi.org/10.20948/abrau-2018-66

10. *Ataeva O.M., Serebryakov V.A., Tuchkova N.P.* Ontological approach to a knowledge graph construction in a semantic library // Lobachevskii J. Math. 2023. Vol. 44, No. 6. P. 2229–2239. https://doi.org/10.1134/s1995080223060471

11. *Ataeva O.M., Tuchkova N.P., Teymurazov K.B. et al.* SciLibRu, the Library of Scientific Subject Domains // Autom. Doc. Math. Linguist. 2025. Vol. 59 (Suppl. 6). P. S505–S512. https://doi.org/10.3103/S000510552570147X

12. *Кобук М.Г., Атаева О.М.* Методы семантической разметки и онтологического моделирования математических текстов в формате LaTeX // Системы высокой доступности. 2026. Т. 22, № 1. С. 90−94. https://doi.org/10.18127/j20729472-202601-18

13. *Khalov A.P., Ataeva O.M., Tuchkova N.P.* Creating a multimodal dataset for the SciLibRu semantic library using a language model // Pattern Recognit. Image Anal. 2026. 36 (In press).

14. *Стребков И.Д.* Метрические инструменты анализа графа знаний предметных областей в семантической библиотеке // Системы высокой доступности. 2026. Т. 22, № 1. С. 95−98. https://doi.org/10.18127/j20729472-202601-19

15. *Халов А.П., Атаева О.М., Тучкова Н.П.* От синтаксиса к семантике: онтология формализации научного знания SciLib // Системы высокой доступности. 2026. Т. 22, № 1. С. 65−70. https://doi.org/10.18127/j20729472-202601-13

16. *Ying H. et al.* Lean Workbook: A large-scale Lean problem set formalized from natural language math problems// arXiv:2406.03847. 2024. https://doi.org/10.48550/arXiv.2406.03847

17. *Peroni S., Shotton D.* The SPAR Ontologies // Proc. 17th Int. Semantic Web Conf. (ISWC 2018). Springer, 2018. P. 119–136.

18. *Brack A. et al.* Requirements Analysis for an Open Research Knowledge Graph // arXiv:2005.10334. arXiv:2005.10334. 2020. https://doi.org/10.48550/arXiv.2005.10334

19. *David C. et al.* Publishing Math Lecture Notes as Linked Data // Proc. CICM 2010. Springer, 2010. P. 370–375.

20. *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMathPRO Ontology: A Linked Data Hub for Mathematics // Communications in Computer and Information Science. Springer, Cham, 2014. Vol. 468. P. 105–119. https://doi.org/10.1007/978-3-319-11716-4_9

21. *Masterman T., Besen, S., Sawtell M., Chao A.* The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey // arXiv:2404.11584. 2024. https://doi.org/10.48550/arXiv.2404.11584

22. *Атаева О.М., Тучкова Н.П.* Методы семантического анализа в процессах обработки данных // Системы высокой доступности. 2026. Т. 22, № 1. С. 99−104. https://doi.org/10.18127/j20729472-202601-20

23. Математическая энциклопедия. В пяти томах. Гл. ред. И. М. Виноградов. М. Советская энциклопедия (1977–1985).

24. *Shoham Y., Leyton-Brown K.* Multiagent systems: algorithmic, game-theoretic, and logical foundations. Cambridge University Press, 2008. 532 p.

25. *Андерсон К.* 73 этапа работы над научным журналом // Научная периодика: проблемы и решения. Т. 5 (23), сентябрь–октябрь 2014. С. 4–10. https://cyberleninka.ru/article/n/73-etapa-raboty-nad-nauchnym-zhurnalom (дата обращения 14.04.2026).

26. *Naddaf M.* AI is transforming peer review – and many scientists are worried// Nature. 2025. Vol. 639. P. 853–854. https://doi.org/10.1038/d41586-025-00894-7

27. *Farber S.* Comparing human and AI expertise in the academic peer re-

view process: towards a hybrid approach // Higher Education Research and Development. 2025. Vol. 44 (4). P. 871–885. https://doi.org/10.1080/07294360.2024.2445575

ORCHESTRATION OF METHODS OF SCIENTIFIC DATA ANALYSIS IN THE REVIEW PROCESSES

O. M. Ataeva[1] [0000-0003-0367-5575], N. P. Tuchkova[2] [0000-0001-5357-9640]

[1, 2]FRC «Computer Science and Control», Russian Academy of Sciences, Moscow, Russia

[1]oataeva@frccsc.ru, [2]ntuchkova@frccsc.ru

Abstract

This paper explores the problem of combining methods in the semantic analysis of scientific data and publications during review. At different stages of data processing in the SciLibRu system, various methods are used, a multi-level ontology is constructed, and a knowledge graph is populated, resulting in the formation of a new data structure distinct from the original. Each method individually serves its purpose in such a system, while their combined use leads to the emergence of new properties, which became the subject of this research. An example of an automatic peer review agent with explainable results is provided.

Keywords: *method orchestration, semantic analysis, domain ontology, knowledge graph, large language models, systems, categories, dynamic structures.*

REFERENCES

1. PRISM. https://openai.com/ru-RU/prism/ (date accessed: 14.04.2026).

2. DoTrace. https://www.domate.ru/dotrace_platform (date accessed: 14.04.2026).

3. *Kubo R.* Thermodynamics: An advanced course with problems and solutions. Amsterdam: North-Holland Publ. Co.; N.Y.: John Wiley and Sons, Inc., 1968. 300 p.

4. A Unified Framework for Self-Organizing Intelligence: A Synthesis of

Computational Autopoiesis, Category Theory, and Iterative Concept-Abstraction Cycles**.** Academia.edu. SSRN. 2025. https://www.academia.edu/143199301/ (date accessed: 14.04.2026).

5. *Tallam K.* From Autonomous Agents to Integrated Systems, A New Paradigm: Orchestrated Distributed Intelligence // arXiv:2503.13754. 2025. https://doi.org/10.48550/arXiv.2503.13754

6. *Bian H.* LLM-empowered knowledge graph construction: A survey // arXiv:2510.20345. 2025. https://doi.org/10.48550/arXiv.2510.20345

7. *Zabihi P., Nawara D., Ibrahim A., Kashef R.* Analyzing Bias in LLM-Augmented Knowledge Graph Systems: Taxonomy, Interaction Mechanisms, and Evaluation // Applied Sciences. 2026. Vol. 16, No. 7. Art. 3410. https://doi.org/10.3390/app16073410

8. *Schintke F. et al.* Validity constraints for data analysis workflows // Future Generation Computer Systems, 2024. Vol. 157. P. 82–97. https://doi.org/10.1016/j.future.2024.03.037

9. *Ataeva O.M., Serebraykov V.A., Tuchkova N.P.* Approaches to the organization of mathematical knowledge when forming subject thesauruses of various mathematics domains // CEUR Workshop Proc. 2018. Vol. 2260. P. 42–54. https://doi.org/10.20948/abrau-2018-66

10. *Ataeva O.M., Serebryakov V.A., Tuchkova N.P.* Ontological approach to a knowledge graph construction in a semantic library // Lobachevskii J. Math. 2023. Vol. 44, No. 6. P. 2229–2239. https://doi.org/10.1134/s1995080223060471

11. *Ataeva O.M., Tuchkova N.P., Teymurazov K.B. et al.* SciLibRu, the Library of Scientific Subject Domains // Autom. Doc. Math. Linguist. 2025. Vol. 59 (Suppl 6). P. S505–S512. https://doi.org/10.3103/S000510552570147X

12. *Kobuk M.G., Ataeva O.M.* Formation of structured representations of scientific journals for integration into a knowledge graph and semantic search // Highly Available Systems. 2026. Vol. 22 (1). P. 90−94 (in Russian). https://doi.org/10.18127/j20729472-202601-18

13. *Khalov A.P., Ataeva O.M., Tuchkova N.P.* Creating a multimodal dataset for the SciLibRu semantic library using a language model // Pattern Recognit. Image Anal. 2026. 36. (In press).

14. *Strebkov I.D.* Metric tools for analyzing the knowledge graph of subject

areas in a semantic library // Highly Available Systems. 2026. 2026. Vol. 22 (1). P. 95−98 (in Russian). https://doi.org/10.18127/j20729472-202601-19

15. *Khalov A.P., Ataeva O.M., Tuchkova N.P.* От синтаксиса к семантике: онтология формализации научного знания SciLib // Highly Available Systems. 2026. Vol. 22 (1). P. 65−70. https://doi.org/10.18127/j20729472-202601-13

16. *Ying H. et al.* Lean Workbook: A large-scale Lean problem set formalized from natural language math problems // arXiv:2406.03847. 2024. https://doi.org/10.48550/arXiv.2406.03847

17. *Peroni S., Shotton D.* The SPAR Ontologies // Proc. 17th Int. Semantic Web Conf. (ISWC 2018). Springer, 2018. P. 119–136.

18. *Brack A. et al.* Requirements Analysis for an Open Research Knowledge Graph // arXiv:2005.10334. 2020. https://doi.org/10.48550/arXiv.2005.10334

19. *David C. et al.* Publishing Math Lecture Notes as Linked Data // Proc. CICM 2010. Springer, 2010. P. 370–375.

20. *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMathPRO Ontology: A Linked Data Hub for Mathematics // Communications in Computer and Information Science. Springer, Cham, 2014. Vol. 468. P. 105–119. https://doi.org/10.1007/978-3-319-11716-4_9

21. *Masterman T., Besen, S., Sawtell M., Chao A.* The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey // arXiv:2404.11584. 2024. https://doi.org/10.48550/arXiv.2404.11584

22. *Ataeva O.M., Tuchkova N.P.* Orchestration of semantic analysis methods // Highly Available Systems. 2026. Vol. 22 (1). P. 99−104. https://doi.org/10.18127/j20729472-202601-20 (in Russian)

23. Matematicheskaya enciklopediya. V 5 tomah. Gl. red. I. M. Vinogradov M. Sovetskaya enciklopediya (1977–1985) (in Russian).

24. *Shoham Y., Leyton-Brown K.* Multiagent systems: algorithmic, game-theoretic, and logical foundations. Cambridge University Press; 2008. 532 p.

25. *Anderson K.* 73 etapa raboty nad nauchnym zhurnalom // Nauchnaya periodika: problemy i resheniya. 2014. T. 5 (23). S. 4–10 (in Russian). https://cyberleninka.ru/article/n/73-etapa-raboty-nad-nauchnym-zhurnalom (date accessed: 14.04.2026)

26. *Naddaf M.* AI is transforming peer review – and many scientists are worried// Nature. 2025. Vol. 639. P. 853–854. https://doi.org/10.1038/d41586-025-00894-7

27. *Farber S.* Comparing human and AI expertise in the academic peer review process: towards a hybrid approach // Higher Education Research and Development. 2025. Vol. 44 (4). P. 871–885. https://doi.org/10.1080/07294360.2024.2445575

СВЕДЕНИЯ ОБ АВТОРАХ

АТАЕВА Ольга Муратовна – старший научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, к. т. н. Область научных интересов: системное программирование, базы данных, инженерия знаний и онтологии.

Olga Muratovna ATAEVA – senior researcher at the Dorodnicyn Computing Center, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences; Candidate of Technical Sciences (PhD). Research interests: systems programming, databases, knowledge engineering, ontologies. Number of scientific publications – 80.

email: oataeva@frccsc.ru

ORCID: 0000-0003-0367-5575

ТУЧКОВА Наталия Павловна – старший научный сотрудник ФИЦ ИУ РАН, кандидат физ.-мат. наук. Специалист в области алгоритмических языков и информационных технологий.

Natalia Pavlovna TUCHKOVA – senior researcher at the Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences (FRC CSC RAS) PhD in Physics and Mathematics. She is an expert in the field of algorithmic languages and information technologies.

email: NTuchkova@frccsc.ru

ORCID: 0000-0001-5357-9640

Материал поступил в редакцию 14 апреля 2026 года