ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2024, том 520, с. 82–89

ИНФОРМАТИКА

УДК 004.82

ОНТОЛОГИИ КАК ФУНДАМЕНТ ФОРМАЛИЗАЦИИ НАУЧНОЙ ИНФОРМАЦИИ И ИЗВЛЕЧЕНИЯ НОВЫХ ЗНАНИЙ

© 2024 г. А. С. Бубнов[1], Н. И. Галлини[2], И. Ю. Гришин[3], И. М. Кобозева[4], Н. В. Лукашевич[5,*], М. Б. Панич[4], Е. Н. Раевский[6], Ф. А. Садковский[4], Р. Р. Тимиргалеева[3]

Представлено академиком РАН А. Л. Семеновым

Получено 20.10.2024 г.
После доработки 08.11.2024 г.
Принято к публикации 08.11.2024 г.

“Ковчег знаний” – цифровой проект, разрабатываемый Московским государственным университетом им. М. В. Ломоносова. Он предоставляет доступ к фундаментальным знаниям на русском языке и должен играть ключевую роль в сохранении и распространении культурного и научного наследия России. “Ковчег знаний” – это онтологическая информационная система. В статье рассматриваются современные представления об онтологии, этапы создания, онтологические особенности БРЭ и Викиданных, а также проектирование информационной системы и применение для обучения языковых моделей. Кратко описан первоначальный рабочий прототип указанной информационной системы. Работы по созданию системы ведутся силами научных сотрудников и программистов лаборатории инженерии знаний Института математических исследований сложных систем МГУ, также учеными филологического, механико-математического факультетов, факультета вычислительной математики и кибернетики, Филиала МГУ в городе Севастополе.

Ключевые слова: онтология, информационная система, фундаментальные знания, проектирование онтологии, информационная система “Ковчег знаний”, Большая российская энциклопедия.

DOI: 10.31857/S2686954324060122, **EDN:** KKGRGT

1. ВВЕДЕНИЕ

В XXI столетии продолжается экспоненциальный рост доступной цифровой информации, частью этого процесса является значительное увеличение количества научных публикаций. Попыткой содействия человеческой деятельности в этих условиях является развитие специализированных информационно-поисковых систем, обеспечивающих удобный доступ к научным публикациям (например, Google Scholar), а также специализированных научных информационных систем, таких как Scopus, Web of Science, eLIBRARY, собирающих качественные источники научной информации и формирующих научные рейтинги [1].

Вместе с тем в настоящее время появились и новые тенденции. Во-первых, в потоке новых публикаций существенно снижается роль статей в авторитетных журналах, поскольку информация начинает распространяться через онлайн сервисы препринтов [2], репозитории моделей и датасетов (например, hugging face [3]), онлайн-лекции и др. [4]. Во-вторых, ситуация с качеством научной информации усложняется в связи с появлением больших языковых моделей, которые могут решать многие задачи автоматической обработки текстов, включая их автоматическое порождение. С одной стороны, такие модели могут помочь исследователю в анализе имеющихся научных публикаций, формулировании собственных результатов, и тем самым ускорять публикацию новых результатов [5]. С другой стороны, такие системы могут автоматически порождать фейковые научные статьи, не имеющие никакой ценности [6–8]. Между этими двумя крайностями есть и широкий спектр проме-

[1] *Лаборатория инженерии знаний Института математических исследований сложных систем, Московский государственный университет им. М. В. Ломоносова, Москва, Россия*

[2] *Крымский федеральный университет им. В. И. Вернадского, Симферополь, Россия*

[3] *Филиал Московского государственного университета им. М. В. Ломоносова в городе Севастополе, Севастополь, Россия*

[4] *Филологический факультет, Московский государственный университет им. М. В. Ломоносова, Москва, Россия*

[5] *Научно-исследовательский вычислительный центр, Московский государственный университет им. М. В. Ломоносова, Москва, Россия*

[6] *Факультет вычислительной математики и кибернетики, Московский государственный университет им. М. В. Ломоносова, Москва, Россия*

[*] *E-mail: louk_nat@mail.ru*

жуточных вариантов, что также представляет собой проблему. Для обучения больших языковых моделей и улучшения результатов их работы как полезных инструментов важно использование качественных данных большого объема [9]. Это особенно существенно для создания специализированных языковых моделей в различных областях науки [10–12].

В связи вышеописанными тенденциями возрастает роль качественного, достоверного научного и образовательного контента, представленного в доверенных онлайн-ресурсах, например, онлайн-энциклопедиях, репозитариях образовательных курсов, специализированных электронных библиотеках и др.

В России в 2016 году соответствии с решениями Правительства Российской Федерации начато создание Большой Российской Энциклопедии (БРЭ) в цифровой форме. Цифровой формат предоставляет значительные возможности для повышения качества, надежности и доступности энциклопедических материалов. Разработка материалов БРЭ является частью более широкой задачи, сохранения, развития, распространения и использования культурного и научного наследия народов России.

В 2023 году Лабораторией инженерии знаний Института математических исследований сложных процессов МГУ были начаты работы по проекту “Ковчег знаний МГУ” (Ковчег). Этот проект можно рассматривать с нескольких сторон:

1) как инкубатор энциклопедических статей: экспертное сообщество Московского университета создает и разрабатывает статьи, которые затем передаются в Большую российскую энциклопедию;

2) как концентратор статей и других информационных источников, создаваемых широким научно-образовательным сообществом России, проходящих рецензирование и редактирование авторитетными представителями научно-образовательного сообщества: заведующими кафедрами вузов, отделами научно-исследовательских организаций, председатели советов по защитам, главные редакторы академических журналов и т. п.;

3) как постоянное хранилище разнообразных информационных источников, таких как учебники, видеозаписи лекций и другие учебные и научные материалы.

Таким образом, Ковчег объединяет в себе как онлайн-энциклопедию, так и научный информационный портал с разнообразными типами информации. Этот проект способствует распространению знаний и обеспечивает доступ к достоверной информации для всех интересующихся. Одним из принципиальных отличий его от информационных собраний типа википедии является использование фильтра в лице авторитетного профессионального сообщества [13].

Создание обширных энциклопедических, научных и образовательных ресурсов требует структуризации хранимых знаний. Традиционно такую функцию играли рубрикаторы, например, рубрикаторы УДК или ГРНТИ, представляющие собой иерархические (древовидные) системы категорий. Цифровые средства позволяют использовать для структуризации более сложно организованные системы, такие как онтологии со специализированными наборами отношений, графы знаний, которые являются семантическими сетями большого объема, содержащими помимо понятий и отношений между ними информацию о конкретных объектах. В случае научной информации такими конкретными объектами являются научные публикации, их авторы, научные институты и др. [4].

Далее мы рассматриваем подходы к онтологической структуризации научной и энциклопедической информации, а также методы построения онтологических ресурсов в рамках Ковчега.

2. ОНТОЛОГИИ И ГРАФЫ ЗНАНИЙ

Онтология — это формализованная система понятий о мире или его фрагменте (предметной области), описывающая объекты и явления, их свойства и взаимосвязи. В отличие от философского понятия онтологии — науки о принципах, видах и основных свойствах бытия, в информатике понятие “онтология” используется для обозначения сети понятий. Эта сеть, также называемая концептуальной или семантической сетью, строится на основе иерархии понятий, связанных отношениями “род — вид” и “вид — экземпляр”. Каждое понятие характеризуется набором свойств и может иметь другие, неонтологические, отношения с другими понятиями. Онтологии играют важную роль в организации знаний, поддерживая структуру наших знаний о мире. Онтология содержит информационный язык для представления знаний, обладающий своим словарем. Он включает единицы определенных типов: понятия (концепты), их экземпляры, имена свойств (параметры и атрибуты), значения параметров и атрибутов, а также имена отношений.

Важным подвидом онтологических ресурсов, используемых для структуризации больших объемов знаний, являются графы знаний. Граф знаний представляет собой семантическую сеть, верхние уровни иерархии которой включают формализованные описания понятий и отношений между ними (собственно, онтологию), а нижние уровни содержат описания конкретных сущностей [14]. Обычно графы знаний содержат представления для сотен тысяч более конкретных сущностей. Графы

знаний стали обсуждаться и развиваться с 2014 года, после их внедрения в интернет-поиск компанией Google [15]. Одним из известных примеров графов знаний в общей предметной области является ресурс Wikidata (Викиданные)[1] [16, 17]. В научной области известным графом знаний является свободно распространяемый ресурс OpenAlex[2], преемник Microsoft Academic Graph (MAG)[3] [4, 18].

Далее мы рассмотрим подходы к структуризации знаний на основе онтологий в энциклопедиях и графах знаний.

3. СТРУКТУРИЗАЦИЯ ИНФОРМАЦИИ В БОЛЬШОЙ РОССИЙСКОЙ ЭНЦИКЛОПЕДИИ

Большая российская энциклопедия (БРЭ), до 1991 года известная как "Большая советская энциклопедия", является универсальной энциклопедией международного уровня и занимает достойное место среди таких справочных изданий, как "Британника", "Ларусс", "Энциклопедия Брокгауз" и др. БРЭ ориентируется на представление целостной картины мира, человека, общества, науки и техники. БРЭ, согласно определению на сайте, представляет собой научно-образовательный портал. Она служит сводом систематизированных достоверных знаний для студентов, преподавателей, ученых, исследователей и широкой аудитории. Основными задачами современного портала БРЭ являются поддержание актуальной электронной национальной базы знаний на русском языке и популяризация науки[4].

В онлайн-версии БРЭ для структуризации статей используется каталог, который представляет собой трехкомпонентную фасетную систему классификации энциклопедических статей. В состав каталога БРЭ входят:

1) области знаний ("Геология", "Искусство" и пр.);

2) категории типов объектов (Астрономы, Горные хребты, Восстания и пр.);

3) теги (ключевые слова).

Каталог БРЭ в 2024 году содержит около 40 областей знаний, более 200 категорий типов объектов и теги, которые можно приписывать к статьям без ограничений. Между областями знаний и категориями нет прямой связи, хотя категории служат для создания своего рода классификации внутри определенной области знаний. Теги — это ключевые слова к энциклопедической статье, они могут быть специфичными, дают возможность навигации по статьям с одними и теми же тегами.

Кроме этого, в энциклопедических статьях объектов разных типов реализовано формализованное описание типовых отношений и атрибутов этих объектов (поля), например, для государств дается информация о параметрах территории и населения, участии в международных организациях и пр. В статьях об исследователях дается информация о годе рождения, государстве, сфере деятельности.

На портале БРЭ реализован поиск по перечисленным компонентам каталога. Каждая статья приписана к некоторой категории типа объекта, которая дает возможность перемещаться по порталу и находить статьи той же категории. При этом навигация по областям знаний по тому же принципу недоступна: приписываемая статье область знаний не является ссылкой. В целом, представленная система дает дополнительные возможности для нахождения в энциклопедии нужной информации. Вместе с тем в системе каталога БРЭ есть особенности, которые усложняют индексацию статей элементами каталога и могут затруднить поиск информации в энциклопедии.

Проведенный нами анализ БРЭ показывает следующее:

1. Категории понятий малочисленны и лишены четкой иерархической структуры. Они не обеспечивают удобную навигацию между статьями.

2. Области знаний организованы непоследовательно и не поддерживают переход по ссылкам для быстрой навигации.

3. Внутритекстовые ссылки являются единственными связями между статьями. Однако большинство ссылок в текущей версии не активно.

4. Теги охватывают различные типы отношений понятий и сущностей, описываемых в статьях. Однако среди тегов много таких, которые относятся только к одной статье, что увеличивает количество тегов и не дает дополнительные возможности для поиска информации. Теги не связаны между собой и могут неявно задавать разные типы отношений.

4. СИСТЕМА ПОНЯТИЙ И ОТНОШЕНИЙ В ВИКИДАННЫХ

"Викиданные", как отдельный проект, представляют собой центральное хранилище структурированных данных, предназначенное для других проектов, например, Википедии, см. [17]. Единица хранения Викиданных имеет имя, описание, позволяющее различать разные единицы с одним именем и любое количество синонимов. Каждая еди-

[1]Wikidata: Q9081. URL: https://www.wikidata.org/wiki/Wikidata:Main_Page (дата обращения: 01.10.2024).

[2]OpenAlex: The open catalogue to the global research system. URL: https://openalex.org/ (дата обращения: 01.10.2024).

[3]Большая российская энциклопедия. О проекте. URL: https://bigenc.ru/p/about-project (дата обращения: 01.10.2024).

[4]Портал. Большая Российская энциклопедия. URL: https://bigenc.ru/ (дата обращения: 01.10.2024).

ница также имеет уникальный идентификатор: Q, за которым следует натуральное число. Информация о единице представлена и доступна человеку и компьютеру на странице единицы в форме нескольких утверждений об этой единице. Утверждения могут в своем составе иметь имена свойств и значения этих имен, например, “место: Германия”. Свойство может иметь несколько значений, таким является например свойство “быть ребенком человека, к которому относится единица”.

Единицы Викиданных связываются классическими онтологическими отношениями (например, “класс-подкласс”, “экземпляр” и др.), к которым добавлено множество новых отношений, работающих по схожему принципу – в Викиданных все они называются свойствами. Например, для объекта Москва (Q649) имеются такие свойства как “instanceof: столица России”, “partof: Центральный федеральный округ”, “foundedby: Юрий Долгорукий” и др. Утверждения также могут содержать квалификаторы, например, указание на время, к которому относится утверждение, источник информации и др. В настоящее время Викиданные представляют собой самую большую открытую базу знаний. 13 марта 2024 года проект “Викиданные” объявил о внесении в базу 2,100,000,000-ой единицы.

Хотя в Викиданных было много усилий уделено аккуратному представлению структурированных данных, тем не менее исследователи обнаруживают достаточно много неточностей и ошибок в описаниях. В частности, есть классы, которые трудно отличить друг от друга, например, *geographical location*, *location*, *geographic region*, *physical location*, *geographical area*. Классы и экземпляры путаются, также смешиваются отношения экземпляр-класс и класс-подкласс. В частности, в работе [17] указано, что для сотен тысяч отношений сообщество авторов Викиданных имеет проблемы с их четким различением. Кроме того, таксономия Викиданных содержит циклы: более чем 47 пар классов являются подклассами друг друга, например *method* и *technique*, имеется 15 циклов длины 3 и больше, например, *axiom*, *first principle*, *principle*.

В работе авторов Викиданных [16] указывается, что в Викиданных была принята гибкая схема для представления знаний, что, с одной стороны, привело к возможности описывать разнообразные виды информации, с другой стороны, привело к проблемам неоднородности и несогласованности данных, которые еще предстоит решить. Несмотря на недостатки, Викиданные используются для создания онтологии Ковчега.

5. СТРУКТУРА ГРАФОВ ЗНАНИЙ

Структуру графов научных знаний мы рассмотрим на примере графа каталога общедоступных ресурсов научных статей, авторов и организаций OpenAlex (название отсылает к Александрийской библиотеке). Каталог является некоммерческим преемником базы Microsoft Academic Graph (MAG). Граф MAG включает шесть типов сущностей: область знаний, автор, организация, статья, место публикации (venue), событие (например, конференция). Граф также включает ссылки цитирования между статьями, для статей указываются авторы, места публикации и области знаний. Данные для графа исходно собирались из электронных библиотек (ACM и IEEE) и с помощью поискового робота системы Bing. Разные копии одних и тех же статей склеиваются в одну сущность графа. Производится процедура автоматического разрешения неоднозначных имен авторов.

Области исследований MAG иерархически организованы в четыре уровня (уровень 0 – 3, где уровень 3 имеет самую высокую гранулярность). Два верхних уровня иерархии областей созданы вручную. На верхнем уровне (уровень 0) находятся 18 областей, выбранных вручную (по алфавиту): биология, бизнес, география, геология, информатика, инженерия, искусство, история, математика, политология, психология, социология, химия, экономика, экология, материаловедение, философия, физика. Области знаний нижних уровней иерархии собираются автоматически с использованием ключевых слов статей.

К марту 2024 года OpenAlex включал метаданные для 209 млн работ, таких как журнальные статьи и книги; 13 миллионов авторов; 124 000 сайтов, на которых размещены работы, включая журналы и онлайн-хранилища; метаданные для 109 000 организаций.

В России некоторым аналогом OpenAlex является информационная система МГУ “Истина” [19, 20], целью которой является сбор, систематизация, хранение, анализ и выдача по запросу информации, характеризующей результаты деятельности научных и образовательных организаций. В настоящий момент система ИСТИНА представляет собой развитую систему научной информации, которая содержит 186857 зарегистрированных пользователей, из которых 30598 работают в МГУ им. М. В. Ломоносова. В частности, в ней содержится информация о 91519 книгах, 392874 докладах на различных научных конференциях, 963917 научных статьях, 136493 учебных курсах.

6. МАКЕТ ИНФОРМАЦИОННОЙ СИСТЕМЫ “КОВЧЕГ ЗНАНИЙ МГУ”

При разработке онтологий, лежащих в основе информационной системы “Ковчег знаний МГУ”, используется комплексный подход, подчеркивающий тесную интеграцию и взаимосвязь как фор-

мальной, так и реально-языковой парадигм. В соответствии с этим, в команду разработчиков этой системы входят специалисты в области прикладной математики, кибернетики, прикладной лингвистики и программисты. Кроме того, неотъемлемой частью процесса являются постоянные консультации с экспертами соответствующих областей знаний (в том числе профессорами и членами российских и международных академий наук).

В качестве формально-математического ориентира при построении системы мы использовали логику определений и понятий [21, 22].

На начальных этапах создания онтологии был использован Protégé – инструмент с открытым исходным кодом, предоставляющий графический интерфейс для разработки, редактирования и визуализации онтологий. Protégé поддерживает стандартный язык веб-онтологий (OWL), а для построения онтологий в области фундаментальных наук в рамках Ковчега мы использовали OWL2 – версию с расширенными функциями и спецификациями по сравнению с ее предшественником.

Создание онтологий высокого уровня является особенно сложной задачей. Такие онтологии объединяют знания, общие для нескольких предметных областей, и позволяют их многократное и многоаспектное использование. Среди наиболее известных крупномасштабных онтологий верхнего уровня – Cyc, DOLCE, SUMO и онтология John Sowa's. Однако общие попытки создать онтологию верхнего уровня, применимую ко всем жизненным сценариям, пока не дали ожидаемых результатов.

В настоящее время идет процесс разработки онтологий для конкретных групп фундаментальных наук (называемых онтологиями 3-го уровня). В нашей работе этот процесс начинается с математики, поскольку большинство других дисциплин полагаются на математические знания в их сегодняшнем состоянии и эволюции развития.

В соответствии с нашей концепцией информационной системы, онтология научных знаний строится в соответствии со следующим требованиями, онтология:

• предлагает гибкие возможности для ввода и организации информации как пользователями так и программными агентами, при этом, предполагается сохранение всех версий и явная фиксация верификации и одобрения в очередной версии, прямой доступ пользователя к последней верифицированной версии;

• обеспечивает унифицированное представление данных, публичную открытость форматов с приоритетом использования стандартизованных, совместимость и возможность интеграции с другими онтологиями, что особенно важно для междисциплинарных исследований; исходно обеспечивается интеграция с системой ИСТИНА МГУ;

• предусматривает различные варианты прав на использование размещаемой информации, с приоритетом открытой лицензии;

• ориентирована на решение задач по анализу, классификации и извлечению информации в ответ на запросы пользователей. В частности, планируется использование системы в системе разведочного научного поиска (exploratory search) [23], предоставляющей пользователю ценные инсайты и рекомендации;

• обеспечивает дружественный интерфейс и техническую поддержку.

Такая система, как Ковчег, способствует формированию целостного научного пространства, где знания не только сохраняются, но и активно обогащаются, постоянно верифицируются, становясь доступными для исследователей и специалистов из смежных областей. При этом существенную роль в создании этой информационной системы играет построение онтологической системы [24]. Указанные процессы и их взаимодействие в процессе создания и совершенствования "Ковчега знаний МГУ" представлены на рис. 1.

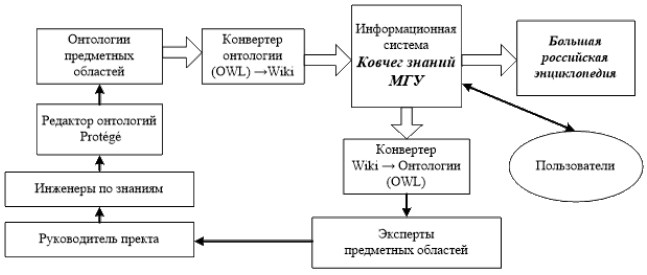

Рис 1. Процесс разработки и совершенствования системы "Ковчег знаний".

В настоящее время разработан и введен в эксплуатацию первый вариант макета указанной информационной системы, который позволяет ученым Московского университета и специалистам наших партнеров принимать участие в наполнении информацией "Ковчега знаний". Пример интерфейса системы приведен на рис. .

Работы по созданию системы ведутся силами научных сотрудников и программистов лаборатории инженерии знаний института математических исследований сложных систем МГУ, возглавляемой академиком РАН А. Л. Семеновым.

7. ЗАКЛЮЧЕНИЕ

В результате проведенных исследований можно сформулировать следующие выводы:

1. Основа системы "Ковчег знаний МГУ": в ходе работы была заложена основа для создания информационной системы. Определены ключевые

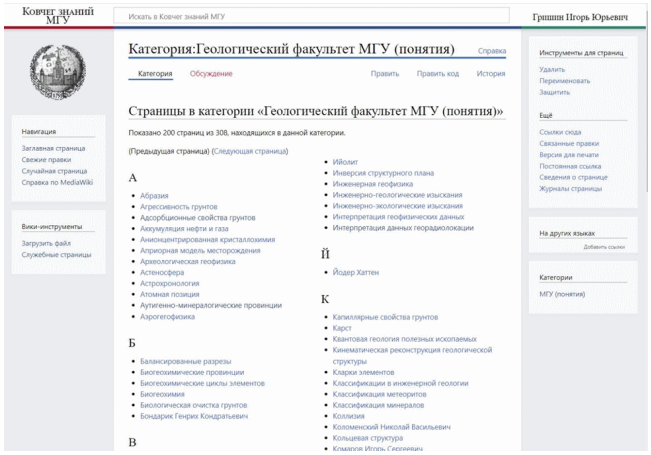

Рис 2. Пример интерфейса информационной системы “Ковчег знаний”.

направления исследований по формированию онтологий в различных естественнонаучных областях. Также разработан первоначальный рабочий прототип информационной системы.

2. Структурирование системы “Ковчег знаний МГУ”: предлагаемый подход к структурированию системы позволяет спроектировать систему, обеспечивающую эффективное хранение, поиск и сортировку научной информации для широкого круга пользователей, включая поддержку авторов БРЭ.

3. Цифровая платформа и совместная работа: система “Ковчег знаний МГУ” построена на цифровой платформе, разработанной МГУ. Первоначальный прототип облегчает совместную работу над базами данных специалистов, участвующих в формировании, разработке и поддержании единой сетевой структуры.

4. Обучение языковых моделей: для обучения языковых моделей используются большие текстовые наборы данных, включая информацию из Википедии. Создание базы данных “Ковчег знаний МГУ” предполагает сбор и систематизацию экспертным сообществом МГУ знаний в различных областях фундаментальной науки. Эта база данных также будет служить обучающим материалом для языковых моделей, способных обрабатывать и интерпретировать сложные научные концепции.

5. Учебник в области научных знаний: создание базы данных “Ковчег знаний МГУ”, формируемой экспертным сообществом Московского университета, может привести к созданию русскоязычного учебника по онтологии фундаментальных научных знаний для высшей школы.

6. Постоянное совершенствование: процесс создания и развития Ковчега носит циклический характер, что позволяет непрерывно обновлять и улучшать онтологии и содержимое системы. Консультации с экспертами и регулярный анализ представленного материала обеспечивают высокое качество предоставляемых данных, что является критически важным для поддержания актуальности системы.

7. Перспективы дальнейшего развития: в будущем планируется расширение функциональности системы, включая интеграцию с внешними источниками данных, такими как базы данных других научных учреждений, что позволит обеспечить более полное и актуальное представление информации. Также рассматривается возможность создания мобильного приложения для удобного доступа к Ковчегу с различных устройств, что повысит доступность системы для широкой аудитории.

8. Влияние на научное сообщество: система “Ковчег знаний МГУ” имеет потенциал для значительного влияния на научное сообщество, облегчая обмен знаниями и содействуя междисциплинарным исследованиям. Создание единой платформы для хранения и обработки научной информации будет способствовать более глубокому пониманию и исследованию сложных проблем, что, в свою очередь, поможет в решении глобальных вызовов.

Мы рассматриваем Проект “Ковчег знаний МГУ” как важный шаг на пути к созданию единой системы для структурирования и хранения научного знания. Он призван не только сохранить ценные данные и результаты исследований, но и облегчить доступ к ним для широкого круга специалистов и исследователей, содействуя новым открытиям и междисциплинарным исследованиям. Благодаря онтологическому подходу к организации и поддержке информации, Ковчег может стать ключевым инструментом для работы с большими объемами научных данных и источником качественных знаний, что особенно актуально в условиях быстрого развития науки и технологий.

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Исследование выполнено при финансовой поддержке Междисциплинарной научно-педагогической школы Московского университета (грант № 23-Щ05-11) и государственного задания (регистрационный № 124020100068-4). Авторы благодарят академика В. А. Садовничего, давшего исходный импульс настоящей работе и постоянно ее поддерживающего.

СПИСОК ЛИТЕРАТУРЫ

1. *Еременко Г. О.* Elibrary.ru: курс на повышение качества контента // Университетская книга, 2016, 3. С. 62–68.
2. *Ginsparg P.* ArXiv at 20 // Nature, 2011, 476(7359). P. 145–147. https://doi.org/10.1038/476145a
3. *Jain S. M.* Introduction to transformers for NLP: With the Hugging Face library and models to solve

problems // Berkeley, CA: Apress, 2022. P. 51–67. ISBN: 9781484288443.

4. *Wang K., Shen Z., Huang C.-Y. et al.* Microsoft academic graph: When experts are not enough // Quantitative Science Studies, 2020, 1(1). P. 396–413. https://doi.org/10.1162/qss_a_00021
5. *Lund B. D., Wang T.* Chatting about ChatGPT: how may AI and GPT impact academia and libraries? // Library hi tech news, 2023, 40(3). P. 26–29. https://doi.org/10.1108/LHTN-01-2023-0009
6. *Haider J., Söderström K. R. Ekström B. et al.* GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation // Harvard Kennedy School Misinformation Review, 2024, 5(5). P. 1–16.
7. *Dadkhah M., Oermann M. H., Hegedüs M. et al.* Detection of fake papers in the era of artificial intelligence // Diagnosis, 2023, 10(4). P. 390–397. https://doi.org/10.1515/dx-2023-0090
8. *Wittau J., Seifert R.* How to fight fake papers: a review on important information sources and steps towards solution of the problem // Naunyn-Schmiedeberg's archives of pharmacology, 2024. P. 1–14. https://doi.org/10.1007/s00210-024-03272-8
9. *Kendall G., da Silva J. A. T.* Risks of abuse of large language models, like ChatGPT, in scientific publishing: Authorship, predatory publishing, and paper mills // Learned Publishing, 2024, 37(1). P. 55–62. https://doi.org/10.1002/leap.1578
10. *Tirumala K., Simig D., Aghajanyan A. et al.* D4: Improving LLM pretraining via document de-duplication and diversification // Advances in Neural Information Processing Systems, 2023, 36. P. 53983–53995. https://doi.org/10.48550/arXiv.2308.12284
11. *Beltagy I., Lo K., Cohen A.* SciBERT: A Pretrained Language Model for Scientific Text // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. P. 3615–3620. https://doi.org/10.18653/v1/D19-1371
12. *Gerasimenko N. A., Chernyavsky A. S., Nikiforova M. A.* RuSciBERT: A transformer language model for obtaining semantic embeddings of scientific texts in Russian // Doklady Mathematics, 2022, 106, Suppl 1. P. S95–S96. https://doi.org/10.1134/S1064562422060072
13. *Горячко В. В., Бубнов А. С., Раевский Е. В., Семенов А. Л.* Цифровой ковчег знаний // Доклады Российской академии наук. Математика, информатика, процессы управления, 2022, 508(1). С. 128–133. https://doi.org/10.31857/S2686954322070098
14. *Hogan A., Blomqvist E., Cochez M, et al.* Knowledge graphs // ACM Computing Surveys (CSUR), 2021, 54(4). P. 1–37. https://doi.org/10.1145/344777
15. *Dong X., Gabrilovich E., Heitz G., et al.* Knowledge vault: A web-scale approach to probabilistic knowledge fusion // Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014. P. 601–610. https://doi.org/10.1145/2623330.2623623
16. *Vrandečić D., Krötzsch M.* Wikidata: a free collaborative knowledgebase // Communications of the ACM, 2014, 57(10). P. 78–85. https://doi.org/10.1145/2629489
17. *Shenoy K., Ilievski F., Daniel Garijo D., et al.* A study of the quality of Wikidata // Journal of Web Semantics, 2022, 72. P. 100679. https://doi.org/10.1016/j.websem.2021.100679
18. *Hug S. E., Ochsner M., Brändle M. P.* Citation analysis with Microsoft academic // Scientometrics, 2017, 111. P. 371–378. https://doi.org/10.1007/s11192-017-2247-8
19. *Васенин В. А. Афонин С. А., Голомазова Д. Д. и др.* Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА) // Информационное общество, 2013, 1–2. С. 39–57.
20. *Козицын А. С., Афонин С. А.* Алгоритм разрешения неоднозначности имен авторов в ИАС ИСТИНА // Современные информационные технологии и ИТ-образование, 2020, 16(1). С. 108–117. https://doi.org/10.25559/SITITO.16.202001.108-117
21. *Семенов А. Л.* Искусственный интеллект в обществе // Доклады РАН. Математика, информатика, процессы управления. Специальный выпуск "Технологии искусственного интеллекта и машинного обучения". 2023, 514(2). С. 6–19. https://doi.org/10.31857/S2686954323350023
22. *Wille R.* Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies // In: Ganter B., Stumme G., Wille R. (eds) Formal Concept Analysis. Lecture Notes in Computer Science, 2005, 3626. Springer, Berlin, Heidelberg. P. 1–33. https://doi.org/10.1007/11528784_1
23. *Лукашевич Н. В., Добров Б. В., Павлов А. М., Штернов С. В.* Онтологические ресурсы и информационно-аналитическая система в предметной области "безопасность" // Онтология проектирования, 2018, 1(27). https://cyberleninka.ru/article/n/ontologicheskie-resursy-i-informionno-analiticheskaya-sis-

tema-v-predmetnoy-oblasti-bezopasnost (дата обращения: 01.10.2024).
24. *Семенов А. Л., Раевский Е. Н., Бубнов А. С. и др.* Универсальная энциклопедическая платформа работы со знанием // Современные информационные технологии и ИТ-образование. 2023, 19(3). С. 696–703. https://doi.org/10.25559/SITITO.019.202303.696-703

ONTOLOGIES AS A FOUNDATION FOR FORMALIZATION OF SCIENTIFIC INFORMATION AND EXTRACTING NEW KNOWLEDGE

A. S. Bubnov[a], N. I. Gallini[b], I. Yu. Grishin[c], I. M. Kobozeva[d], N. V. Lukashevich[e], M. B. Panich[c], E. N. Raevsky[f], F. A. Sadkovsky[c], R. R. Timirgaleeva[c]

[a] *Knowledge Engineering Laboratory, Institute for Mathematical Research of Complex Systems, Lomonosov Moscow State University, Moscow, Russia*
[b] *Vernadsky Crimean Federal University, Simferopol, Russia*
[c] *Branch of Lomonosov Moscow State University in the city of Sevastopol, Sevastopol, Russia*
[d] *Faculty of Philology, Lomonosov Moscow State University, Moscow, Russia*
[e] *Research Computing Center, Lomonosov Moscow State University, Moscow, Russia*
[f] *Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia*

Presented by the Academician of the RAS A. L. Semenov

"Ark of Knowledge" is a digital project developed by M. V. Lomonosov Moscow State University. It provides access to fundamental knowledge in Russian and should play a key role in the preservation and dissemination of Russia's cultural and scientific heritage. "Ark of Knowledge" is an ontological information system. The article discusses modern ideas about ontology, stages of creation, ontological features of BDT and Wikidata, as well as the design of an information system and the use of language models for training. The initial working prototype of this information system is briefly described. Work on creating the system is being carried out by researchers and programmers from the Knowledge Engineering Laboratory of the Institute for Mathematical Research of Complex Systems of Moscow State University, as well as scientists from the Faculty of Philology, Mechanics and Mathematics, the Faculty of Computational Mathematics and Cybernetics, and the Branch of Moscow State University in Sevastopol.

Keywords: ontology, information system, fundamental knowledge, ontology design, information system "Ark of Knowledge", Great Russian Encyclopedia.