УДК 004.62+004.22 *Научная статья* DOI: 10.18287/2223-9537-2026-16-2-341-354

Построение графа знаний предметной области на основе открытых электронных словарей

© 2026, Н.А. Евстифеева ✉, И.А. Ширеторова

Национальный исследовательский технологический университет «МИСиС» (МИСиС), Москва, Россия

Аннотация

Рассматривается построение графа знаний с целью объединения знаний по химии, физике и физической химии в области водородной энергетики и технологических процессов получения водорода. Разработанный граф знаний содержит формализацию семантических связей и абстрактную организацию классов, сопоставимую с онтологиями предметной области. Описана система поддержки принятия решений для расширения графа знаний на основе алгоритма многокритериальной оценки валидации внешних источников данных в виде открытых электронных словарей. При валидации источника терминологических данных предметной области в системе поддержки принятия решений используется Интернет-сервис стандартизованной оценки. Собранные в интерфейсе оценочные средства автоматически предоставляют данные для экспертного заключения с принятием решения и собирают дополнительные данные для нейронной сети, которая обучается верификации внешних источников на прецедентах с экспертным решением. Граф знаний содержит 23,5 тысячи уникальных сущностей, представленных на русском языке с дублированным значением на английском языке.

Ключевые слова: *граф знаний, онтологическая модель, предметная область, электронный словарь, многокритериальная оценка, открытые сервисы, валидация онтологии.*

Цитирование: *Евстифеева Н.А., Ширеторова И.А.* Построение графа знаний предметной области на основе открытых электронных словарей. *Онтология проектирования*. 2026. Т.16, №2(60). С.341-354. DOI: 10.18287/2223-9537-2026-16-2-341-354.

Вклад авторов: *Евстифеева Н.А.* – разработка графа знаний; формализация обратной связи и ситуационного управления; *Ширеторова И.А*. – выбор внешних сервисов для оценки; разработка программного комплекса.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Современные большие языковые модели (БЯМ) имеют склонность к галлюцинациям. Для устранения галлюцинаций в ответах БЯМ используются мультиагентный подход или технологии, связанные с дополнительной оценкой и принятием решения (ПР) [1]. Для увеличения глубины знаний БЯМ в предметной области (ПрО) часто применяются генерация с расширением через поиск (*Retrieval Augmented Generation, RAG*-архитектура) и выравнивание с применением метода адаптации предварительно обученной модели [2].

Онтологические модели (ОМ) ПрО используются для формирования границ информационных полей ПрО, покрывающих большие области знаний. Это позволяет формализовать разрабатываемую модель знаний как граф знаний (ГЗ) в мультиграфовой объектно-ориентированной структуре. *Целью данной статьи* является разработка структурированной модели знаний ПрО, ограниченной физическими и химическими процессами, технологиями и инженерными решениями для технологического процесса получения водорода. Для этого необходимо построить алгоритм автоматического мониторинга информационного пространства ПрО с автоматизацией оценки и последующего ПР по валидации внешнего источника данных с целью расширения разработанного ГЗ.

Декомпозиция поставленной цели привела к решению двух задач: разработка специализированной структуры ГЗ и первичное её наполнение на основе подобранного экспертами корпуса текстов; создание средств для расширения ГЗ за счёт знаний из тематически близких открытых Интернет-источников, которые можно типизировать как ОМ ПрО. Для решения задачи расширения разработанной структуры необходима система многокритериальной оценки и валидации источника в виде электронного словаря (ЭлС).

1 Онтологическая модель в формализме графа знаний

В компьютерной реализации структуры текстовой информации ПрО при построении ОМ часто используется набор представительских примитивов [3]. В данной статье под представительскими примитивами понимаются: классы и множество их экземпляров; атрибуты, являющиеся свойствами классов и их экземпляров; отношения, характеризующие связи между экземплярами классов. Словари ПрО формально причисляются к онтологиям, но не все словари могут считаться ОМ ПрО [4].

В Интернете доступны открытые источники в виде ЭлС по различным ПрО в форме глоссариев, тезаурусов, семантических словарей и др. Они отличаются по структуре данных и по составу содержащейся информации. Для компьютерного представления ОМ принято использовать *Resource Description Framework* (*RDF*), который согласуется с *Ontology Web Language* (*OWL*) [5]. Применимость открытых источников из Интернета для автоматизации процесса расширения ГЗ как ОМ ПрО зависит от их состава и структуры и требует дополнительной оценки с ПР [6, 7].

Онтологии делятся на три класса: онтологии верхнего уровня, онтологии ПрО и прикладные онтологии [8]. Большинство известных онтологий в рассматриваемой ПрО либо узкоспециализированы, либо формализованы из базовых объектов ПрО и не имеют междисциплинарных связей, характерных для водородной энергетики [9].

Разработанная модель представляет собой гибридную структуру, сочетающую свойства ГЗ и ОМ, предназначена для структурирования и объединения междисциплинарных знаний из разнородных источников [10].

На рисунке 1 представлена диаграмма классов верхнего уровня разработанного ГЗ.

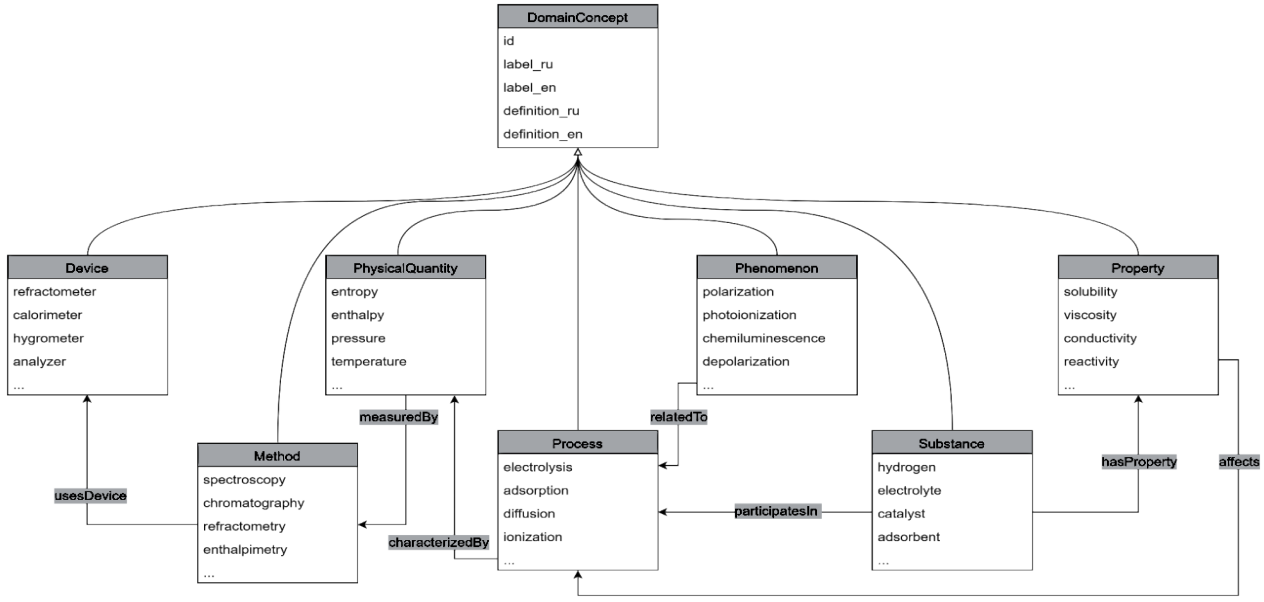


Рисунок 1 – Диаграмма классов верхнего уровня разработанного графа знаний

Базовый класс *DomainConcept* задаёт общие атрибуты понятий ПрО, включая идентификаторы, наименования и определения. Производные классы *Substance, Process, Property, PhysicalQuantity, Method, Device* и *Phenomenon* формируют семантическую структуру ГЗ, которая включает отношения, соответствующие предикатам: *hasProperty, participatesIn, characterizedBy, measuredBy, usesDevice, relatedTo, affects*.

Все вершины в разработанном ГЗ представляют собой фреймы, которые приведены к стандарту представления данных *OWL* и включают классы *owl*:*Class* либо экземпляры классов *owl*:*NamedIndividual*. Атрибуты вершин ГЗ представляют собой слоты фреймов, типизированные по свойствам и методам, и в рамках *OWL* им сопоставлены: *owl*:*DatatypeProperty* для типа слотов свойства из ГЗ; слоты, отражающие функциональные и семантические связи в ГЗ, которые соответствуют объектным свойствам *owl*:*ObjectProperty*. Триплетное представление знаний в разработанном ГЗ вида *subject-predicate-object* (*SPO*) соответствует структуре *RDF/OWL*, где субъект и объект интерпретируются как индивидуумы или классы, а предикат – как объектное свойство, задающее отношение между ними. Иерархическая структура классов отображается в виде отношений наследования *rdfs*:*subClassOf*.

2 Сравнение графа знаний с открытыми электронными словарями

Применение машинного перевода с учётом специфики ПрО в рамках БЯМ показывает высокое качество результатов [10]. Для оценки качества автоматического перевода с английского языка на русский сформирована контрольная выборка из 100 терминов, случайным образом отобранных из химических и физико-химических групп ГЗ. Экспертная проверка показала, что 84% переводов являются полностью корректными, 11% – частично корректными и требуют редакторской правки, а 5% – некорректными. Это позволяет использовать машинный перевод как рабочий инструмент первичного пополнения при расширении ГЗ в автоматическом режиме из англоязычных открытых источников.

В качестве примера приведены ОМ из области химии, которые представляют собой ЭлС на английском языке. Они покрывают разные аспекты ПрО и поддерживаются проектными экспертными группами в ручном режиме обновления.

- *Chemical Entities of Biological Interest* (*ChEBI*) является онтологией в виде тезауруса, активно используемого с 2008 года [11], содержит описание малых молекул и биологически значимых веществ.
- *Chemical Functional Ontology* (*ChemFOnt*) – это иерархическая онтология, содержащая функции и роли химических веществ. Она содержит более 341000 химических веществ и 515000 терминов и определений, структурированных в четыре базовых функциональных аспекта, 12 суперкатегорий и более 173700 ветвлений внутри иерархии [12].
- *Chemical Methods Ontology* (*CHMO*) является ОМ, ориентированной на описание аналитических методов, например методов масс-спектрометрии, хроматографии и др. [13]. Термины построены на основе *IUPAC Orange Book*[1]. Онтология имеет более 3000 классов.

Количественная оценка перекрытия терминологии между открытыми онтологиями и разработанным ГЗ проводилась на основе сопоставления векторных представлений терминов, сформированных с использованием многоязычной нейросетевой модели (НСМ) *paraphrase-multilingual-MiniLM-L12-v2*. С её помощью получены семантические представления пар на русском и английском языках для терминов. Векторизация выполнялась по нормализованным и лематизированным текстовым представлениям пар терминов, а текстовые определения использовались как дополнительный контекст при экспертной проверке пограничных случаев. Векторное представление формировалось по всей фразе целиком без усреднения отдельных слов, что позволяло учитывать контекстное значение. В качестве меры близости использовалась косинусная метрика, а термин считался покрытым при превышении

[1] Gold Book. Compendium of Chemical Terminology. Version 5.0.0. (14588 Terms). https://goldbook.iupac.org/.

порогового значения сходства 0.7. Результаты показали, что степень пересечения реализованного на первичном этапе ГЗ с рассматриваемыми онтологиями остаётся ограниченной и существенно зависит от их тематической направленности. Так, для онтологии *CHMO* общее покрытие составляет 2–3%. Онтология *ChemFOnt* показала более высокий уровень соответствия: *Structure* ~24%; *DomainConcept* ~13%; *Substance* ~12%. Однако общее покрытие полного словаря не превысило ~6%.

Проведённое сравнение показывает, что существующие онтологии могут быть использованы как внешние источники терминологии и частично – структурные ориентиры, однако не могут быть применены в качестве основы для ГЗ в ПрО, связанной с производством водорода. Необходимо расширить наполнение разработанного ГЗ знаниями разнородных типов, включая химические и физические закономерности, математические расчёты и инженерные решения, формализацию межсущностных связей в рамках единой семантической модели.

3 Оценка выбора источника данных для расширения графа знаний

На первом этапе расширения ГЗ как ОМ ПрО необходимо произвести оценку и принять решения по валидации ЭлС как источника данных. Разработанный алгоритм основан на ситуационной модели управления с применением обратной связи, как управляющего воздействия в случае принятия допустимости применения источника данных. При валидации источника производится предварительная коррекция данных, и верифицированный ЭлС приводится к стандартизованной модели с учётом особенностей разработанного ГЗ.

ГЗ имеет в качестве источников не только ЭлС, но и корпус текстов, подобранных экспертами, в виде документов, обеспечивающих расширенную информацию о ПрО. Схема программного комплекса представлена на рисунке 2. Корпус текстов подобран экспертной группой по заданной ПрО для автоматического анализа с целью построения ГЗ как ОМ по принципу триплетной формализации с представлением предиката в виде функциональной связи первого уровня.

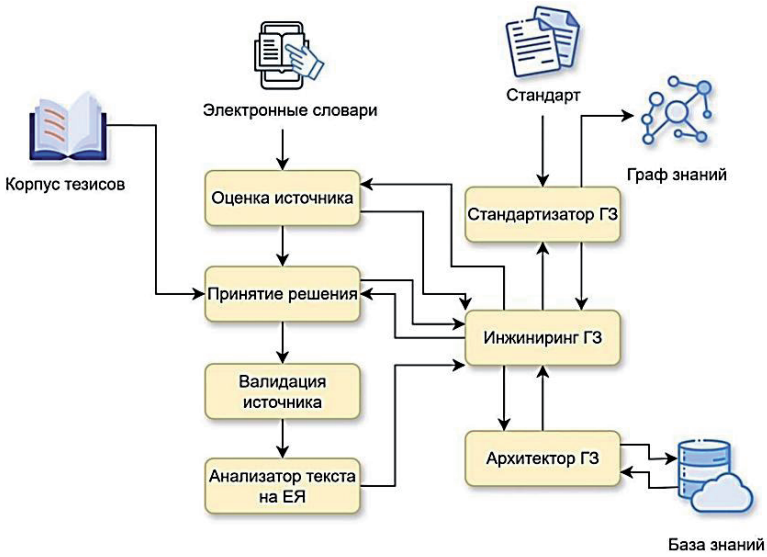


Рисунок 2 – Схема разработанного программного комплекса

В таком наполнении разработанного ГЗ данные ЭлС формируют базовую терминологическую основу классов и их экземпляров как ОМ ПрО, а корпус текстов увеличивает полноту междисциплинарного пересечения областей знаний. Модули, представленные на рисунке 2, позволяют проводить добавление источников текстовых данных с последующей реструктуризацией ГЗ. Валидация источников в форме ЭлС выполняется в два этапа: автоматическая оценка модулем с системой количественных критериев; автоматизированное экспертное решение с использованием системы поддержки и ПР.

Модуль многокритериальной оценки источников содержит два алгоритма, исполняемых в параллельных потоках: расчёт многокритериальной оценки; проверка на внешних сервисах по стандартизованным критериям. Внешние веб-сервисы для оценки источников подобраны

в соответствии с международным стандартом ISO 25964 [14, 15], в котором указаны правила создания и обмена словарями и тезаурусами, включая мультиязычные и совместимые по требованиям *Simple Knowledge Organization System* (*SKOS*). Задачей *SKOS* является обеспечение представления тезаурусов и других видов онтологий в Интернете.

Сервис *Ontology Evaluation Framework* использован для оценки онтологий на непротиворечивость, полноту, когерентность и гибкость. С помощью *Ontology Recommender* 2.0 можно по заданным параметрам, используемым как ключевые слова для поиска, подобрать онтологии с последующей оценкой по метрикам покрытия и валидации.

Для многокритериальной оценки источников разработаны пять групп критериев с индивидуальным весом для каждого критерия в группе:

- *надёжность*: открытость источника, *API, FAIR*-принципы (*Findable, Accessible, Interoperable, Reusable* – находимость, доступность, совместимость и пригодность);
- *масштабируемость*: совместимость с графовой моделью и *SPO*;
- *непротиворечивость*: уникальность, абсолютная частотная индивидуальная характеристика, относительная частотная парная характеристика, относительная частотная индивидуальная характеристика, синонимия, омонимия;
- *полнота*: объём классов, объём экземпляров, объём связей, полнота связанности, полнота элементов связи, полнота фрейма сущности;
- *формальная совместимость*: соответствие стандартам *RDF, OWL* и *SKOS*.

Для расчёта критерия соответствия *FAIR*-принципам в разработанном алгоритме используется Интернет-сервис *FAIR Ontology Testing* (*FOOPS*!) [16]. Оценки возможности использования Интернет-источника для расширения ГЗ, полученные с применением *FOOPS*!, позволяют верифицировать корректность используемых источников и служат встроенным в общий интерфейс разработанной системы инструментом для экспертной оценки.

По указанным группам критериев автоматически вычисляется поэтапная свёртка в соответствии с эмпирически подобранными весовыми коэффициентами значимости критериев. В процессе свёртки значений частных критериев используется взвешенная линейная комбинация, позволяющая получить интегральную оценку валидности источника, как сумму нормированных критериев с учётом их весовых коэффициентов. Последовательность выполнения многокритериальной оценки внешнего источника представлена на рисунке 3.

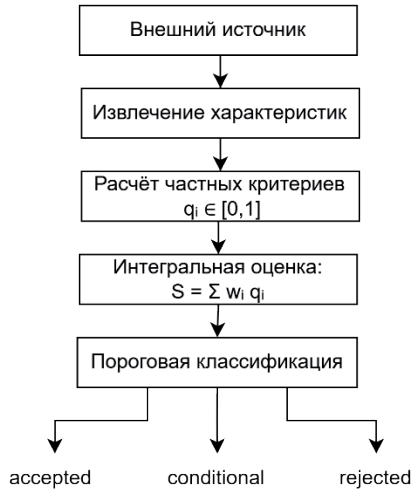


Рисунок 3 – Схема алгоритма многокритериальной оценки внешнего источника

Подбор весовых коэффициентов выполнялся на основе экспертной калибровки по контрольной выборке внешних источников данных. В оценке участвовали три эксперта, которые оценивали значимость пяти групп критериев для задачи расширения ГЗ по пятибалльной шкале с последующим усреднением и нормировкой результатов. В результате получены следующие значения весов: надёжность 0.30, полнота 0.25, непротиворечивость 0.20, масштабируемость 0.15 и формальная совместимость 0.10. Более высокие значения коэффициентов для надёжности и полноты обусловлены критической важностью достоверности источника и достаточности терминологической информации при расширении ГЗ.

Для проверки согласованности экспертных оценок использовался коэффициент Кендалла. При проведении анализа чувствительности веса варьировались в пределах 10% с последующим пересчётом итоговых оценок. Анализ показал, что для большинства источников итоговый класс сохраняется («условно принятый» – *conditional*, «принятый» – *accepted* и «отклонённый» – *rejected*). На основе разметки источников обучена НСМ, позволяющая определять допустимость использования ЭлС для расширения ГЗ.

4 Использование нейросетевой модели в условиях неоднозначности

Алгоритм ПР на основе НСМ повышает устойчивость системы в условиях неоднозначности. На вход модели подаётся вектор признаков, сформированный на основе агрегированных значений пяти групп критериев. Интегральный показатель в состав входных признаков не включается, что позволяет избежать влияния целевой переменной и обеспечивает корректность обучения модели. Архитектура НСМ включает два скрытых слоя размерности 16 и 8 нейронов с функцией активации *ReLU*. На выходе формируется вероятностная оценка принадлежности источника к одному из классов: принятый, условно принятый, отклонённый.

Обучение НСМ проводилось на выборке из 360 источников, размеченных экспертами. При формировании обучающей выборки экспертами учтено частичное пересечение значений критериев между классами, что позволило смоделировать реальные условия ПР. В качестве целевой переменной использовалась экспертная оценка допустимости интеграции источника. Для сопоставления использовалась базовая модель, основанная на пороговой классификации по интегральной оценке. Качество модели оценивалось с использованием стандартных метрик многоклассовой классификации. При разделении выборки в соотношении 80/20 тестовая часть составила 72 источника из 360 размеченных объектов. Базовая модель показала *accuracy* ≈ 0.82 и *F1-score* ≈ 0.81, тогда как НСМ достигла *accuracy* ≈ 0.89 и *F1-score* ≈ 0.89. Это показывает, что НСМ более устойчиво обрабатывает пограничные случаи, в которых базовая модель недостаточно учитывает их совместное влияние при линейной свёртке критериев. НСМ показала хорошую степень согласованности с экспертными решениями при сохранении способности к обобщению.

В качестве примера экспертной оценки внешнего источника с использованием разработанной системы ПР получены значения критериев надёжности 0.29, полноты 0.81, согласованности 1.00, масштабируемости 0.93 и формальной совместимости 0.78; итоговое значение критерия допустимости интеграции источника в ГЗ составило 0.732.

Анализ результатов валидации Интернет-источников показал, что НСМ позволяет корректировать решения, полученные на основе критериальной оценки. Для источников с высокой полнотой и надёжностью, но частично выраженной формальной совместимостью, линейная модель даёт заниженную итоговую оценку, относя такие источники к классу условно принятых; НСМ классифицировала их как принятые, учитывая совокупное влияние признаков. Для источников с неоднородной структурой НСМ в ряде случаев понижала оценку по сравнению с пороговым правилом, что позволяет избежать ложноположительных решений.

На рисунке 4 представлена матрица ошибок модели классификации. Анализ матрицы ошибок показывает, что НСМ демонстрирует высокую точность классификации, что выражается в преобладании значений на главной диагонали.

Функция потерь, представленная на рисунке 5, показывает устойчивую сходимость модели в процессе обучения, что свидетельствует о корректной настройке параметров модели и отсутствии выраженного переобучения.

Комбинация многокритериальной оценки и обучаемой НСМ обеспечивает баланс между интерпретируемостью и адаптивностью принимаемых решений, а полученные результаты показывают работоспособность предложенного подхода.

Для оценки применимости разработанного ГЗ проведён эксперимент по сравнению ответов БЯМ в двух режимах: без использования ГЗ; с использованием структурированного контекста, извлекаемого из ГЗ. Тестовая выборка формировалась автоматически на основе терминов и определений, входящих в разработанный ГЗ. Для каждого элемента из тестовой выборки извлекалось определение термина и выполнялась автоматическая оценка семантического соответствия эталонному определению с выявлением признаков галлюцинаций.

Оценка качества ответов проводилась автоматически на основе косинусного сходства векторных представлений, проверки появления отсутствующих в ГЗ терминов и числовых значений. На рисунке 6 представлены результаты сравнения двух режимов.

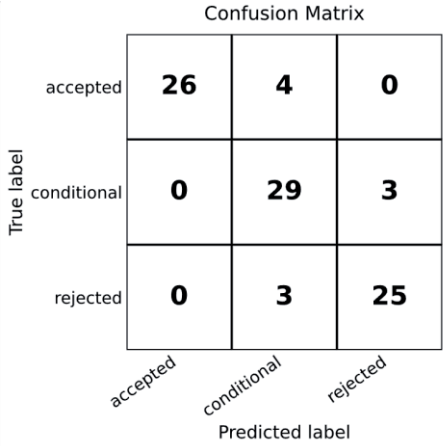


Рисунок 4 – Матрица ошибок нейросетевой модели классификации источников

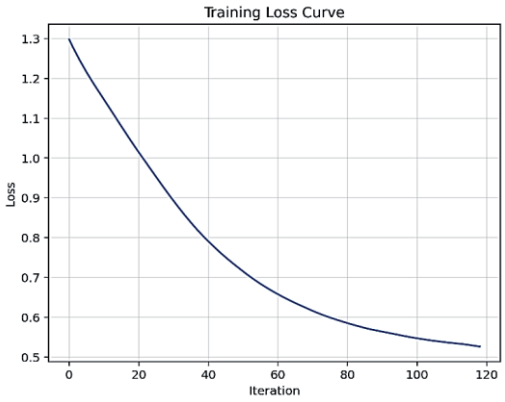
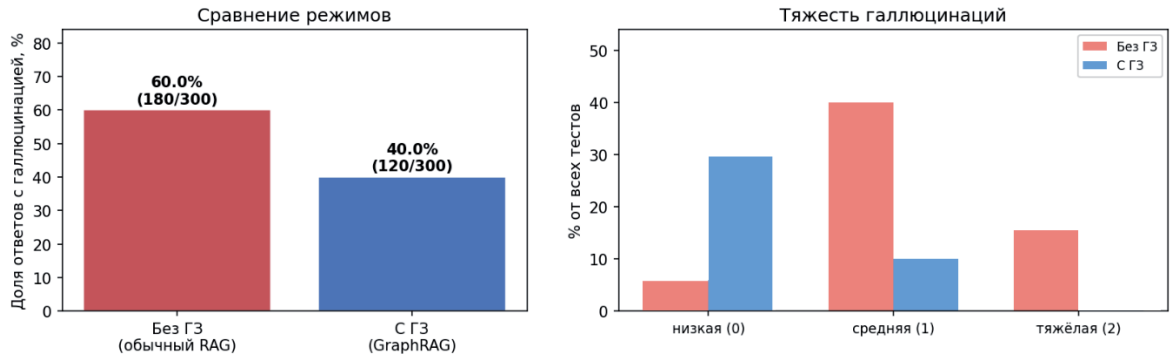


Рисунок 5 – Кривая функции потерь в процессе обучения нейросетевой модели



а) б)

Рисунок 6 – Сравнение доли (а) и тяжести галлюцинаций (б) большой языковой модели при использовании графа знаний

Полученные результаты подтверждают, что структурированный ГЗ может использоваться как внешний источник достоверного контекста для уменьшения генерации недостоверной информации БЯМ.

Трудности возникли при агрегировании данных ЭлС как между собой, так и с ГЗ, построенным по корпусу текстов. Это связано со сложностью обобщения вершин итогового ГЗ с формализацией нескольких типов связей для любой вершины, а также фактическим участием вершин в семантически разнотипных предикатах. Такая многозначность необходима для получения полного описания возможных сценариев технологических процессов в ПрО и требует подбора алгоритмов разрешения неоднозначностей при структуризации данных для максимальной сохранности представленных знаний [17, 18].

5 Расширение разработанного графа знаний

ГЗ является ориентированным, взвешенным и непланарным мультиграфом. Рёбрами ГЗ являются предикаты из триплетного представления знаний. В качестве предикатов для сохранения подробных семантических связей определяются глагольные грамматические конструкции, выражающие функциональные связи в ГЗ. Такая структура является базовым ГЗ

(БГЗ), на основе которого с применением алгоритма интеллектуальной обработки строятся иерархические связи между вершинами – экземплярами классов, формализованными в итоговом ГЗ. Вершины БГЗ классифицируются на два типа фреймов: объекты и субъекты. Каждый фрейм вершины разработанного ГЗ сформирован из слотов со специальной информацией и двух типов семантически значимых слотов: свойства и метод. Свойства – сущности, которые не отнесены в анализируемом предложении к субъекту/объекту, но характеризуют субъект либо объект, например, прилагательными или числовыми значениями. К методам относятся функциональные связи второго уровня, которые лингвистически представляют собой причастный или деепричастный обороты. При этом в итоговом БГЗ в слотах типа методы присутствуют только сами причастия или деепричастия, а все другие составные члены причастного или деепричастного оборота трансформируются в слоты с типом свойство.

Так как выделение триплетов из корпуса текстовой информации выполняется автоматически и сопровождается пополнением уже реализованного ГЗ, на этапе извлечения могут формироваться неуникальные триплеты. Поэтому первоначальный набор триплетов рассматривается как мультимножество, в котором один и тот же триплет может иметь несколько вхождений. При построении БГЗ выполняется переход к множеству уникальных триплетов с сохранением информации о частоте их появления: $Tr_{unique} = \{ tr_i \mid freq(tr_i) \geq 1 \}$, где $tr_i = (s_i, p_i, o_i)$ – триплет, включающий субъект, предикат и объект, а $freq(tr_i)$ – число его вхождений в исходное мультимножество. Совокупность субъектов образует множество S, совокупность предикатов – множество P, а совокупность объектов – множество O. Множество вершин БГЗ формируется как объединение субъектов и объектов: $V = S \cup O$, предикаты задают рёбра между ними. Такое представление позволяет одновременно устранить дублирующиеся триплеты при построении структуры графа и сохранить частотные характеристики их появления в модели данных.

Формализация мультимножества, состоящего из триплетов, сохраняется в реляционной модели данных как трёхмерная матрица из отношений частоты встречаемости сочетаний субъекта, предиката и объекта в моделируемой ПрО. Данная матрица используется при анализе с целью выделения иерархических связей и пересчитывается в относительную частоту по типу векторной модели *TF-IDF* [19].

Для полноты и однозначности представления сущностей в БГЗ необходимо провести оценку с установлением матрицы соответствия между вершинами БГЗ и слотами фреймов, типизированными как свойства. В результате работы алгоритма данные остаются во множественном представлении, что увеличивает эффективность анализа для выявления признакового пространства вершин графа [20, 21].

Слоты, типизированные как свойства и методы, объединяются в рамках одного уровня иерархии, вошедших в выделенный класс, и автоматически проверяются на синонимичность и близость по метрике косинусного сходства в рамках векторной модели. Значения слотов атрибутов, признанные синонимичными, формализуются как одно значение и позиционируются как атрибут реляционного отношения. Так как реляционное отношение не может иметь повторяющихся атрибутов, класс получает свойства и методы без повторений [22, 23].

Наследование от нижнего уровня иерархии к верхнему обеспечивается расширением реляционного отношения на каждом уровне с сохранением всех атрибутов предыдущих уровней, определённых как свойства и методы классов-наследников. Процесс обновления структуры классов в части свойств и методов включает анализ выявления свойств и методов, уже имеющихся в классе, и добавляемых: однотипных, идентичных, синонимов и антонимов.

Основным источником химических терминов, использованных при расширении ГЗ, является пятая редакция *IUPAC Gold Book*, опубликованная Международным союзом чистой и

прикладной химии и содержащая 12380 терминов на английском языке. Пример добавленного в ГЗ термина показан на рисунке 7.


```
[
  {
    "id": "f7afaeb1d0574d2ae2eaac8ec14af3ccea18cb2e3d481935383efaa081d7cb33",
    "topic_id": "a5b231852fa5e23b140828f83253901ce3f4a4915a118c3e3a0c430b6e544c06",
    "raw_text": "AM 0 солнечный свет",
    "cleaned_text": null,
    "lemmatized_text": "am 0 солнечн свет",
    "first_letter": "a",
    "language": "ru",
    "info": "Auto-generated from dataset",
    "created_at": "2025-07-16T17:10:20.052454",
    "descriptions": [
      {
        "id": "fc6e0068cc9833f7c3e065713b2db8287de308a87950f03e6d4e459b3e1ac910",
        "word_id": "f7afaeb1d0574d2ae2eaac8ec14af3ccea18cb2e3d481935383efaa081d7cb33",
        "raw_text": "Солнечное излучение в космосе непосредственно над атмосферой Земли
        "cleaned_text": null,
        "lemmatized_text": "солнечн излучен космос непосредствен атмосфер земл плоскост
        "language": "ru",
        "info": null,
        "created_at": "2025-07-16T17:10:20.091815",
        "embeddings": []
      }
    ],
    "triplets": []
```


Рисунок 7 – Пример добавленного в граф знаний термина из *IUPAC Gold Book*

Каждая запись вершины ГЗ представлена в виде структурированного объекта со следующими атрибутами:

id – уникальный идентификатор сущности (вершины) в ГЗ;
topic_id – идентификатор тематического кластера (класса), к которому отнесена данная сущность в иерархии ГЗ;
raw_text – текстовое представление термина, извлечённое из источника;
cleaned_text – очищенное текстовое представление;
lemmatized_text – лемматизированное представление термина, используемое для унификации и сопоставления сущностей;
first_letter – первая буква термина, используемая для индексирования и оптимизации поиска;
language – язык термина;
info – служебная информация о происхождении записи (например, источник);
created_at – временная метка создания записи;
descriptions – список описаний сущности, представленных в виде вложенных объектов.

Каждый элемент списка *descriptions* включает:

id – уникальный идентификатор описания;
word_id – ссылка на идентификатор сущности, к которой относится описание;
raw_text – исходный текст описания;
cleaned_text – очищенное представление текста описания;
lemmatized_text – лемматизированное описание;
language – язык описания;
info – дополнительная служебная информация;
created_at – временная метка создания описания;
embeddings – векторное представление описания, используемое для задач семантического анализа.

Поскольку *JSON*-структура содержит не только содержательные элементы сущности, но и служебные поля, то элементы сопоставляются как с классами разработанного ГЗ, так и с атрибутами и свойствами *RDF/OWL*.

Для расширения терминологической базы выбраны пять валидированных источника, относящихся к трём областям знаний:

- *для химической области*: 5-ая редакция *IUPAC Gold Book*, опубликованная Международным союзом чистой и прикладной химии и содержащая 12380 терминов на английском языке;
- *для физической области*: физико-математический словарь-справочник (содержит 4500 статей на русском языке и переводом на английский и немецкий языки) [24]; англо-русский словарь по физике плазмы и управляемому термоядерному синтезу (содержит около 3 тыс. англоязычных терминов с русским толкованием) [25];
- *для физико-химической области*: *Physical Chemistry* (содержит около 650 англоязычных терминов по термодинамике, химическому равновесию, кинетике, квантовой химии и т.д.); *Atkins' Physical Chemistry* (включает около 1000 терминов по термодинамике, химической кинетике, молекулярной структуре, спектроскопии и статистической термодинамике).

Формализация специальных данных системного характера, позволяет обеспечить связанность структур ГЗ с источниками текстовой информации, которые хранятся как физические объекты и в виде ссылок. В связи с большой размерностью, хранилище подвергается дополнительному архивированию и структуризации с предварительной обработкой для сокращённого представления текстовых фрагментов или структур данных в быстром доступе и ссылками на полное представление объектов в архиве.

Оценка расширения ГЗ внешними источниками приведена в таблице 1. Из таблицы видно значительное увеличение числа вершин и рёбер.

Таблица 1 – Количественная оценка графа знаний до и после расширения

Показатель	До расширения	После расширения	Прирост
Число вершин	2 327	23 532	+21 205
Число рёбер	6 814	68 120	+61 306
Плотность графа	0.0025	0.012	+0.0095

В таблице 2 указано число сущностей для каждого источника, использованного при расширении ГЗ. Наибольший вклад в расширение ГЗ обеспечил ресурс *IUPAC Gold Book*. Такое распределение подтверждает, что химический компонент остаётся ядром данной ПрО, а физика и физическая химия обеспечивают уточнение междисциплинарных связей.

Таблица 2 – Показатели расширения графа знаний по различным источникам для трёх предметных областей

Источник	Область	Добавлено сущностей
IUPAC Gold Book	химия	12380
Физико-математический словарь + словарь по физике плазмы	физика	5792
Physical Chemistry Glossary + Atkins' Physical Chemistry	физическая химия	3001

Выводы

Модель информационной системы разработана на основе модульной архитектуры и представляет собой масштабируемую платформу для подключения в качестве компонента *RAG*-архитектурного решения выравнивания БЯМ без дообучения в границах ПрО. Одним из основных модулей является комплекс ПР, который обеспечивает возможность многоуровневого оценивания внешних электронных ресурсов как ОМ ПрО по признакам применимости для автоматического расширения ГЗ ПрО водородных технологий. Расширение ГЗ внешними ЭлС повысило качество ответов БЯМ за счёт структурирования характеристик и параметров, выделения синонимических групп и иерархических связей ОМ ПрО.

- Разработанная модель данных предусматривает логико-семантические взаимосвязи между понятиями химии, физики, физической химии, инженерных технологий и обеспечивает совместимость по *RDF/OWL* с существующими онтологиями и открытыми словарями, автоматизируя категоризацию сущностей. ГЗ ориентирован на междисциплинарную интеграцию и включает связи между различными типами сущностей (например: «вещество участвует в процессе», «процесс характеризуется физической величиной», «метод использует устройство»).
- Система ПР обеспечивает автоматизацию сбора данных с последующим обучением НСМ по выбору наиболее валидного внешнего источника текстовой информации. Оценка качества НСМ составила значение 0,89.
- Из автоматически проанализированных предобученной НСМ 12 внешних источников терминологических данных пять были признаны валидными для интеграции в ГЗ. По экспертной оценке правильности отклонения семи источников были выявлены их несоответствия по критериям структурированности, полноты или совместимости с разработанной моделью ГЗ.
- Качество машинного перевода зависит от объёма переводимых токенов для каждого элемента ГЗ и составило 84–95%. Созданный ГЗ имеет более чем 23,5 тыс. сущностей на английском и русском языках в ПрО.
- В ходе экспериментальной оценки выявлен ряд ограничений связанных с качеством расширения ГЗ, которое зависит от полноты, непротиворечивости, однозначности и структуры внешних источников данных, а также от качества машинного перевода. Дополнительную сложность представляют омонимия терминов и различия терминологических традиций при интеграции междисциплинарных источников. При объединении большого количества ЭлС возрастает риск накопления ошибок сопоставления, дублирования сущностей и появления противоречивых определений. Требуется дополнительная адаптации используемой НСМ при переходе к другим ПрО.

Список источников

[1] ***Jones C.R., Bergen B.K.*** Large language models pass the Turing test [Электронный ресурс]. — arXiv:2503.23674. DOI: 10.48550/arXiv.2503.23674.

[2] ***Сидорова Е.А., Иванов А.И., Овчинникова К.А.*** Извлечение информации из текстов на основе онтологии и больших языковых моделей. *Онтология проектирования*. 2025. Т.15, №1(55). С.114–129. DOI: 10.18287/2223-9537-2025-15-1-114-129.

[3] ***Darwish A.M., Rashed E.A., Khoriba G.*** Mitigating LLM hallucinations using a multi-agent framework. *Information*. 2025. Vol.16, No.7. Art.517. DOI: 10.3390/info16070517.

[4] ***Beatty A.S., Kaplan R.M.*** **(eds.)**. Ontologies in the behavioral sciences: Accelerating research and the spread of knowledge. Washington, DC: National Academies Press, 2022. 164 p. DOI: 10.17226/26464.

[5] ***Booshehri M. et al.*** Introducing the Open Energy Ontology: Enhancing data interpretation and interfacing in energy systems analysis. *Energy and AI*. 2021. Vol.5. Art.100074. DOI: 10.1016/j.egyai.2021.100074.

[6] ***Antoniou G., van Harmelen F.*** The Semantic Web Primer. Cambridge: MIT Press, 2005. 238 p.

[7] ***Smith A.*** Simple Knowledge Organization System (SKOS). *Knowledge Organization*. 2022. Vol.49, No.5. P.371–384. DOI: 10.5771/0943-7444-2022-5-371.

[8] ***An B. et al.*** Accurate text-enhanced knowledge graph representation learning. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL-HLT 2018). New Orleans, 2018. P.745–755. DOI: 10.18653/v1/N18-1068.

[9] ***Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьёв В.Д.*** Онтологии и тезаурусы: модели, инструменты, приложения. М.: Бином. Лаборатория знаний, 2009. 176 с.

[10] ***Головин А.А., Жукова Н.А.*** Построение графа знаний по телекоммуникационным данным. *Онтология проектирования*. 2025. Т.15, №1(55). С.45–54. DOI: 10.18287/2223-9537-2025-15-1-45-54.

[11] ***Küçük D.*** OntoWind: An improved and extended wind energy ontology [Электронный ресурс]. arXiv:1803.02808. DOI: 10.48550/arXiv.1803.02808.

[12] ***Degtyarenko K. et al.*** ChEBI: An Open Bioinformatics and Cheminformatics Resource. *Current Protocols in Bioinformatics*. 2009. Vol.14. P.14.9.1–14.9.20. DOI: 10.1002/0471250953.bi1409s26.
[13] ***Wishart D.S. et al.*** ChemFOnt: The chemical functional ontology resource. *Nucleic Acids Research*. 2023. Vol.51, No.D1. P.D1220–D1229. DOI: 10.1093/nar/gkac919.
[14] ***Strömert P. et al.*** Ontologies4Chem: The landscape of ontologies in chemistry // *Pure and Applied Chemistry*. — 2022. Vol.94, No.6. P.605–622. DOI: 10.1515/pac-2021-2007.
[15] **ISO 25964-2:2013.** Information and documentation – Thesauri and interoperability with other vocabularies – Part 2: Interoperability with other vocabularies. https://www.iso.org/standard/53658.html.
[16] ***Alexiev V., Isaac A., Lindenthal J.*** On the composition of ISO 25964 hierarchical relations (BTG, BTP, BTI). *International Journal on Digital Libraries*. 2016. Vol.17. P.39–48.
[17] **FOOPS!** FAIR ontology testing. https://w3id.org/foops/. Ontology Pitfall Scanner for FAIR (Beta) https://foops.linkeddata.es/FAIR_validator.html.
[18] ***Paulheim H.*** Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*. 2017. Vol.8, No.3. P.489–508. DOI: 10.3233/SW-160218.
[19] ***Nickel M. et al.*** A review of relational machine learning for knowledge graphs [Электронный ресурс]. arXiv:1503.00759. DOI: 10.48550/arXiv.1503.00759.
[20] ***Parhomenko P.A., Grigorev A.A., Astrakhantsev N.A.*** A survey and an experimental comparison of methods for text clustering: Application to scientific articles. *Proceedings of the Institute for System Programming of RAS*. 2017. Vol.29, No.2. P.161–200. DOI: 10.15514/ISPRAS-2017-29(2)-6.
[21] ***Dorodnykh N., Yurin A.*** Knowledge graph engineering based on semantic annotation of tables. *Computation*. 2023. Vol.11, No.9. Art.175. DOI: 10.3390/computation11090175.
[22] ***Tiddi I., Schlobach S.*** Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*. 2022. Vol.302. Art.103627. DOI: 10.1016/j.artint.2021.103627.
[23] ***Gao T., Yao X., Chen D.*** SimCSE: Simple contrastive learning of sentence embeddings. *Proceedings of EMNLP* 2021. P.6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552.
[24] ***Алленицын А.Г., Бутиков Е.И.*** Физико-математический словарь-справочник. Том 2: Англо-русский словарь. СПб.: Политехника, 2011. 437 с.
[25] ***Курнаев В.А., Курнаев А.А.*** Англо – русский толковый словарь по физике плазмы и управляемому термоядерному синтезу. М.: НИЯУ «МИФИ». 2014. 101 с.

Сведения об авторах

Евстифеева Наталья Александровна, 1977 г. рождения. Окончила Московский институт Стали и сплавов (2001), к.т.н. (2005). Доцент кафедры инженерной кибернетики НИТУ МИСиС. В списке научных трудов работы в области САПР и ИИ. Author ID (SPIN-код): 2305-5942; Author ID (Scopus): 59014004200; ORCID: 0009-0001-9079-366X. *evstifeeva@mail.ru.* ✉.

Ширеторова Ирина Артуровна, 2003 г. рождения. Окончила бакалавриат НИТУ МИСиС (2024). Магистрант кафедры инженерной кибернетики МИСиС. ORCID: 0009-0001-2161-4496. *irinashir902@gmail.com*.

Поступила в редакцию 15.01.2026, после рецензирования 29.05.2026. Принята к публикации 2.06.2026.

Scientific article DOI: 10.18287/2223-9537-2026-16-2-341-354

Construction of a domain knowledge graph based on open electronic dictionaries

© 2026, N.A. Evstifeeva ✉, I.A. Shiretorova

National University of Science and Technology MISIS, Moscow, Russia

Abstract

This paper addresses the construction of a knowledge graph aimed at integrating knowledge from chemistry, physics, and physical chemistry in the field of hydrogen energy and hydrogen production processes. The developed knowledge graph incorporates formalized semantic relationships and an abstract class organization comparable to domain ontologies. A decision support system for knowledge graph expansion is described. The system is based on a multicriteria evaluation algorithm for validating external data sources represented by open electronic dictionaries. During the validation of a domain-specific terminological data source, the decision support system employs a standardized online assessment service. The evaluation tools integrated into the interface automatically provide data for expert review and decision-making, while simultaneously collecting additional data for a neural network trained to verify external sources using cases with expert-approved decisions. The resulting knowledge graph contains 23,500 unique entities represented in Russian and supplemented with corresponding English-language entries.

Keywords: *knowledge graph, ontological model, domain, electronic dictionary, multicriteria evaluation, open services, ontology validation.*

For citation: *Evstifeeva N.A., Shiretorova I.A.* Construction of a domain knowledge graph based on open electronic dictionaries [In Russian]. *Ontology of designing.* 2026; 16(2)*:* 341-354. DOI: 10.18287/2223-9537-2026-16-2-341-354.

Authors' contributions: *Evstifeeva N.A.* – knowledge graph development; formalization of feedback and situational management*; Shiretorova I.A.* – selection of external evaluation services; development of the software package.

Conflict of interest: The authors declare no conflict of interest.

List of figures and tables

Figure 1 – Upper-level class diagram of the developed knowledge graph
Figure 2 – Architecture of the developed software system
Figure 3 – Workflow of the multicriteria evaluation algorithm for external sources
Figure 4 – Confusion matrix of the neural network model for source classification
Figure 5 – Loss function curve during neural network training
Figure 6 – Comparison of the proportion (a) and severity (b) of hallucinations generated by a large language model when using the knowledge graph
Figure 7 – Example of a term added to the knowledge graph from the IUPAC Gold Book
Table 1 – Quantitative evaluation of the knowledge graph before and after expansion
Table 2 – Knowledge graph expansion indicators for various sources across three subject domains

References

[1] ***Jones CR, Bergen BK.*** Large language models pass the Turing test. arXiv:2503.23674. DOI: 10.48550/arXiv.2503.23674.

[2] ***Sidorova EA, Ivanov AI, Ovchinnikova KA.*** Information extraction from texts based on ontology and large language models [In Russian]. *Ontology of designing*. 2025; 15(1): 114–129. DOI: 10.18287/2223-9537-2025-15-1-114-129.

[3] ***Darwish AM, Rashed EA, Khoriba G.*** Mitigating LLM hallucinations using a multi-agent framework. *Information*. 2025; 16(7): Art. 517. DOI: 10.3390/info16070517.

[4] ***Beatty AS, Kaplan RM. (eds.)***. Ontologies in the behavioral sciences: Accelerating research and the spread of knowledge. Washington, DC: National Academies Press, 2022. 164 p. DOI: 10.17226/26464.

[5] ***Booshehri M. et al.*** Introducing the Open Energy Ontology: Enhancing data interpretation and interfacing in energy systems analysis. *Energy and AI*. 2021; 5: Art. 100074. DOI: 10.1016/j.egyai.2021.100074.
[6] ***Antoniou G, van Harmelen F.*** The Semantic Web Primer. Cambridge: MIT Press, 2005. 238 p.
[7] ***Smith A.*** Simple Knowledge Organization System (SKOS). *Knowledge Organization*. 2022; 49(5): 371–384. DOI: 10.5771/0943-7444-2022-5-371.
[8] ***An B. et al.*** Accurate text-enhanced knowledge graph representation learning // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL-HLT 2018). — New Orleans, 2018. P.745–755. DOI: 10.18653/v1/N18-1068.
[9] ***Dobrov BV, Ivanov VV, Lukashevich NV, Solovyov VD.*** Ontologies and thesauri: models, tools, applications [In Russian]. Moscow: Binom, 2009. 176 p.
[10] ***Golovin AA, Zhukova NA.*** Building a knowledge graph from telecommunication data [In Russian]. *Ontology of designing*. 2025; 15(1): 45–54. DOI: 10.18287/2223-9537-2025-15-1-45-54.
[11] ***Küçük D.*** OntoWind: An improved and extended wind energy ontology. arXiv:1803.02808. DOI: 10.48550/arXiv.1803.02808.
[12] ***Degtyarenko K. et al.*** ChEBI: An Open Bioinformatics and Cheminformatics Resource. *Current Protocols in Bioinformatics*. 2009; 14: 14.9.1–14.9.20. DOI: 10.1002/0471250953.bi1409s26.
[13] ***Wishart D.S. et al.*** ChemFOnt: The chemical functional ontology resource. *Nucleic Acids Research*. 2023; 51(D1): D1220–D1229. DOI: 10.1093/nar/gkac919.
[14] **Strömert P. et al.** Ontologies4Chem: The landscape of ontologies in chemistry // *Pure and Applied Chemistry*. — 2022. — Vol. 94, No. 6. — P. 605–622. — DOI: 10.1515/pac-2021-2007.
[15] **ISO 25964-2:2013.** Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies. https://www.iso.org/standard/53658.html.
[16] ***Alexiev V, Isaac A, Lindenthal J.*** On the composition of ISO 25964 hierarchical relations (BTG, BTP, BTI). *International Journal on Digital Libraries*. 2016; 17: 39–48.
[17] ***FOOPS!*** FAIR ontology testing. https://w3id.org/foops/. Ontology Pitfall Scanner for FAIR (Beta) https://foops.linkeddata.es/FAIR_validator.html.
[18] ***Paulheim H.*** Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*. 2017; 8(3): 489–508. DOI: 10.3233/SW-160218.
[19] ***Nickel M. et al.*** A review of relational machine learning for knowledge graphs. arXiv:1503.00759. DOI: 10.48550/arXiv.1503.00759.
[20] ***Parhomenko PA, Grigorev AA, Astrakhantsev NA.*** A survey and an experimental comparison of methods for text clustering: Application to scientific articles. *Proceedings of the Institute for System Programming of RAS*. 2017; 29(2): 161–200. DOI: 10.15514/ISPRAS-2017-29(2)-6.
[21] ***Dorodnykh N, Yurin A.*** Knowledge graph engineering based on semantic annotation of tables. *Computation*. 2023; 11(9): Art. 175. DOI: 10.3390/computation11090175.
[22] ***Tiddi I, Schlobach S.*** Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*. 2022; 302: Art. 103627. DOI: 10.1016/j.artint.2021.103627.
[23] ***Gao T, Yao X, Chen D.*** SimCSE: Simple contrastive learning of sentence embeddings. *Proceedings of EMNLP*. 2021. P.6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552.
[24] ***Alenitsyn AG, Butikov EI.*** Physical and mathematical reference dictionary. Vol. 2: English–Russian dictionary. [In Russian]. Saint Petersburg: Polytechnic, 2011. 437 p.
[25] ***Kurnaev VA, Kurnaev AA.*** English-Russian explanatory dictionary on plasma physics and controlled thermonuclear fusion [In Russian]. Moscow: NRNU MEPhI, 2014. 101 p.

About the authors

Natalia Aleksandrovna Evstifeeva (b. 1977) graduated from the Moscow Institute of Steel and Alloys (MISIS) in 2001, Candidate of Technical Sciences (2005). Associate Professor at the Department of Engineering Cybernetics of the National University of Science and Technology MISIS. Author of scientific publications in the fields of CAD systems and AI. Author ID (RSCI): 2305-5942; Scopus Author ID: 59014004200; ORCID: 0009-0001-9079-366X. *evstifeeva@mail.ru.* ✉

Irina Arturovna Shiretorova (b. 2003) graduated from the National University of Science and Technology MISIS with a Bachelor's degree in 2024. Master's student at the Department of Engineering Cybernetics of MISIS. ORCID: 0009-0001-2161-4496. *irinashir902@gmail.com.*

Received January 15, 2026. Revised May 29, 2026. Accepted June 2, 2026.