ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATION TECHNOLOGIES

УДК 004.65;004.75;004.5;004.91
https://doi.org/10.37661/1816-0301-2024-21-2-7-23

Оригинальная статья
Original Article

Основы функционирования семантического портала ядерных знаний BelNET

С. Н. Сытова✉, В. В. Гавриловец, А. П. Дунец, А. Н. Коваленко, С. В. Черепица

Институт ядерных проблем
Белорусского государственного университета,
ул. Бобруйская, 11, Минск, 220006, Беларусь
✉*E-mail: sytova@inp.bsu.by*

Аннотация

Цели. Рассмотрена возможность использования семантических технологий для развития и совершенствования системы управления контентом научно-образовательного портала eLab-Science и созданного на ее основе белорусского портала ядерных знаний BelNET (Belarusian Nuclear Education and Training Portal, https://belnet.by/).

Методы. Разработаны оригинальные алгоритмы автоматической систематизации – размещения записей контента в таксономии портала на основе семантических технологий и формирования списка ключевых слов. Использованы такие понятия семантических технологий, как таксономия (иерархическая структура портала), тезаурус, глоссарий.

Результаты. Разработанные алгоритмы реализованы и протестированы с использованием инструмента полнотекстового поиска и оригинального белорусского глоссария по ядерной и радиационной безопасности.

Заключение. Описанные принципы организации и алгоритмы на базе семантических технологий, лежащие в основе функционирования системы управления контентом научно-образовательного портала eLab-Science и созданного на ее базе белорусского портала ядерных знаний BelNET, позволяют эффективно реализовывать размещение записей контента в таксономии портала, а также автоматически формировать набор ключевых слов создаваемого ресурса.

Ключевые слова: ядерные знания, управление ядерными знаниями, информационная система, свободное программное обеспечение, тезаурус, глоссарий, таксономия

Благодарности. Работа выполняется в рамках мероприятия 13 «Выполнение работ по оказанию научно-технической поддержки Министерству по чрезвычайным ситуациям Республики Беларусь в области обеспечения ядерной и радиационной безопасности» подпрограммы 3 «Научное обеспечение эффективной и безопасной работы Белорусской атомной электростанции и перспективных направлений развития атомной энергетики» Государственной программы «Наукоемкие технологии и техника» на 2021–2025 гг.

Для цитирования. Основы функционирования семантического портала ядерных знаний BelNET / С. Н. Сытова [и др.] // Информатика. – 2024. − Т. 21, № 2. – С. 7–23.
https://doi.org/10.37661/1816-0301-2024-21-2-7-23

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 05.12.2023
Подписана в печать | Accepted 21.03.2024
Опубликована | Published 28.06.2024

Basics of the semantic portal of nuclear knowledge BelNET functioning

Svetlana N. Sytova✉, Viktar V. Haurylavets, Andrei P. Dunets, Anton N. Kavalenka, Siarhei V. Charapitsa

Institute for Nuclear Problems
of Belarusian State University,
st. Bobruiskaya, 11, Minsk, 220006, Belarus
✉*E-mail: sytova@inp.bsu.by*

Abstract

Objectives. The possibility of using semantic technologies for the development and improvement of the content management system of the scientific and educational portal eLab-Science and the Belarusian nuclear knowledge portal BelNET (Belarusian Nuclear Education and Training Portal, https://belnet.by/) created on its basis is being considered.

Methods. Original algorithms for automatic systematization have been developed, such as placing content records in the portal taxonomy based on semantic technologies and generating a list of keywords. The following concepts of semantic technologies are used: taxonomy (hierarchical structure of the portal), thesaurus, glossary.

Results. The developed algorithms were implemented and tested using a full-text search tool and the original Belarusian glossary on nuclear and radiation safety.

Conclusion. The described basic principles of organization and algorithms based on semantic technologies, which underlie the functioning of the content management system of the scientific and educational portal eLab-Science and the Belarusian nuclear knowledge portal BelNET, created on its basis, make it possible to effectively implement the placement of content records in the portal taxonomy, as well as automatically generate a set of keywords for the resource being created.

Keywords: nuclear knowledge, nuclear knowledge management, information system, free software, thesaurus, glossary, taxonomy

Acknowledgements. The work is carried out within the framework of the activity 13 "Performing work to provide scientific and technical support to the Ministry of Emergency Situations of the Republic of Belarus in the field of ensuring nuclear and radiation safety" of Subprogram 3 "Scientific support for the effective and safe operation of the Belarusian nuclear power plant and promising directions for the development of nuclear energy" of the State Program "High-tech technologies and equipment" for 2021–2025.

For citation. Sytova S. N., Haurylavets V. V., Dunets A. P., Kavalenka A. N., Charapitsa S. V. *Basics of the semantic portal of nuclear knowledge BelNET functioning*. Informatika [*Informatics*], 2024, vol. 21, no. 2, pp. 7−23 (In Russ.). https://doi.org/10.37661/1816-0301-2024-21-2-7-23

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

Введение. Ядерные знания [1] связаны как с исследованиями и разработками, так и с промышленным использованием ядерных технологий и включают широкий спектр энергетических и неэнергетических применений. Международное агентство по атомной энергии (МАГАТЭ) с начала 2000-х гг. активно разрабатывает методологию и руководящие документы для планирования, разработки и реализации программ управления ядерными знаниями [2].

Согласно терминологии МАГАТЭ [3] управление ядерными знаниями: получение, сбор, передача, сохранение, поддержание и использование знаний, а также обмен ими – имеет важное значение для развития и поддержания технических знаний и компетенций, необходимых для ядерно-энергетических программ и различных ядерных технологий. Деятельность МАГАТЭ в этой области содействует ядерному образованию, предоставляя поддержку, возможности для налаживания связей и обмена опытом [4]. В последние годы одним из наиболее действенных инструментов в менеджменте ядерных знаний являются порталы ядерных знаний, создаваемые и развиваемые при поддержке МАГАТЭ, а также разнообразные корпоративные системы современных ядерных знаний [5].

В соответствии с материалами МАГАТЭ [6] семантическая технология, лежащая в основе веб-поиска и управления онлайн-информацией, в обязательном порядке должна использоваться в ядерной области для помощи экспертам и всем заинтересованным сторонам в поддержании, сохранении и обмене ядерными знаниями. Применение семантических технологий помогает в интеграции различных источников данных, автоматизации индексации ресурсов, облегчает поиск информации эффективным и экономичным способом, повышает устойчивость управления сложными и междисциплинарными системами ядерной энергетики. В настоящее время МАГАТЭ развивает различные инициативы в области семантических технологий, которые могут принести пользу в области менеджмента ядерных знаний. Это касается как проектов МАГАТЭ, так и национальных программ [6].

В качестве яркого примера можно привести созданную в 1970 г. под эгидой МАГАТЭ Международную ядерную информационную систему INIS (The International Nuclear Information System, https://www.iaea.org/resources/databases/inis), используемую более чем в 130 странах мира. Репозиторий INIS содержит библиографические ссылки, научные и технические отчеты, материалы конференций, патенты и тезисы во всех областях деятельности МАГАТЭ, включая ядерную технику и технологии, ядерную безопасность и радиационную защиту, гарантии и нераспространение, применение ядерных и изотопных методов, ядерную физику и физику высоких энергий, ядерную и радиационную химию, ядерные применения в науках о жизни, правовые аспекты, экологические и экономические аспекты ядерных и неядерных источников энергии. При предметной классификации (категоризации) каждая запись в базах данных INIS отнесена к определенной предметной категории, а также к одной или нескольким вторичным тематическим категориям.

В Беларуси в настоящее время формируется полноценная система управления ядерными знаниями, основу которой составляет портал ядерных знаний BelNET [7–10]. Его цели полностью соответствуют подходам МАГАТЭ к менеджменту ядерных знаний.

Статья посвящена становлению активно развивающегося в настоящее время портала ядерных знаний BelNET как семантического портала. Необходимость использования семантических технологий в оригинальной системе управления контентом портала BelNET на основе свободного программного обеспечения вызвана тем, что ранее разработчики столкнулись с отрицательным опытом при ручной систематизации (определении разделов и подразделов портала, в которых должна быть размещена запись) создаваемых на портале BelNET новых записей. В результате этого многие разделы таксономии (иерархической структуры портала), которая оказалась очень большой и сложной, до сих пор остаются пустыми либо слабозаполненными. Однако очевидно, что многие ресурсы могли бы быть размещены в нескольких разделах и стать более доступными для читателей.

Теоретические основы семантических технологий. Семантическая технология является одним из бурно развивающихся алгоритмических направлений современной прикладной математики и информационных технологий. Она включает в себя широкий спектр инструментов, стандартов и методологий, позволяющих обрабатывать информацию в зависимости от ее контекста и значения. Назовем основные понятия семантических технологий, используемые в работе: онтология, глоссарий, тезаурус, таксономия.

Онтология используется для подробной формализации области знаний с помощью концептуальной схемы, которая состоит из структуры данных, содержащей все релевантные классы объектов, их связей и правил (теоремы, ограничения), принятых в этой области. Основные

сферы применения онтологий – моделирование бизнес-процессов, семантическая паутина (https://www.w3.org/standards/semanticweb/) и искусственный интеллект. Данные и онтология с правилами вывода вместе представляют собой базу знаний [11] предметной области.

Глоссарий – это словарь узкоспециализированных терминов в отрасли знаний с толкованием, иногда переводом на другой язык, комментариями и примерами. Он не использует дополнительные связи между терминами и может рассматриваться как онтология с пустым множеством отношений.

Тезаурус – это словарь с дополнительными отношениями, охватывающий понятия, определения и термины области знаний или сферы деятельности, которые подчиняются семантическим отношениям между терминами. Обычные простейшие таксономические отношения в тезаурусах составляют несколько уровней отношений типа выше-ниже [12]. Специализированный тезаурус может разрабатываться экспертами или строиться с помощью программных средств [13]. Для создания тезаурусов существуют специальные государственные и международные стандарты[1, 2]. На основе тезауруса может быть создана таксономия (иерархическая структура) портала. Отметим правило [14], что если в тезаурусе нет подходящего дескриптора для поиска полезного понятия, то следует предложить и ввести в тезаурус новый.

В качестве примера тезауруса приведем разработанный МАГАТЭ многоязычный тезаурус для системы INIS (https://inis.iaea.org/search/thesaurus.aspx) на арабском, китайском, английском, французском, немецком, японском, русском и испанском языках, предоставляющий переводы тысяч технических терминов, которые помогают в навигации и поиске по коллекции INIS. Тезаурус дает возможность пользователям БД INIS индексировать и искать литературу на нескольких языках. Объем английской версии тезауруса [15] в настоящий момент составляет 31 301 термин.

Для онлайн-поиска в поисковых запросах могут использоваться контролируемые термины (дескрипторы) – ключевые слова, произвольные текстовые слова или их комбинация, предназначенные для предметного индексирования с контролируемой терминологией по заголовку ресурса и свободному тексту (например, аннотации, реферату, полному тексту ресурса). Для единообразия такие дескрипторы должны входить в тезаурус или глоссарии. При выборе релевантных ссылок из результатов поиска очень полезными элементами являются реферат, заголовок и дескрипторы исследуемого ресурса.

Таксономия – это иерархическая структура портала, которая может быть построена на основании семантических технологий, в частности на основании одного или нескольких тезаурусов [14].

Оригинальный тезаурус портала BelNET. При работе над таксономией портала BelNET было принято решение придерживаться комбинированного подхода с использованием наработок МАГАТЭ и большого багажа знаний белорусских экспертов.

Для внедрения семантических технологий на портале ядерных знаний BelNET разработан тезаурус (дерево категорий). Его верхний уровень изображен на рис. 1. Здесь находятся разделы «Глоссарий по ядерной и радиационной безопасности», «Научные глоссарии», «Организации», «География», «Персоналии», «Календарь и события». Объем этих глоссариев не должен быть большим. При необходимости внесения нового термина глоссарии всегда могут быть дополнены. Оптимальный объем тезауруса в настоящий момент составляет примерно две-три тысячи терминов.

Глоссарий по ядерной и радиационной безопасности включает 525 терминов. Он специально разработан для портала BelNET на основе нескольких глоссариев МАГАТЭ, Госкорпорации «Росатом», НАТО, а также белорусских нормативно-правовых документов в области ядерной и радиационной безопасности. Глоссарий предназначен для обеспечения алгоритмов функцио-

[1]Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления : ГОСТ 7.25-2001. – Введ. 01.07.02. – М. : Изд-во стандартов, 2001. − 14 с.

[2]Информация и документация. Тезаурусы и взаимосвязь с другими словарями. Ч. 1. Тезаурусы для выдачи информации : ИСО 25964-1:2011. – Введ. 08.08.11. – М. : Изд-во стандартов, 2011. − 160 с.

нирования портала ядерных знаний, а также облегчения понимания и использования основных терминов в области ядерной и радиационной безопасности с учетом белорусской специфики.

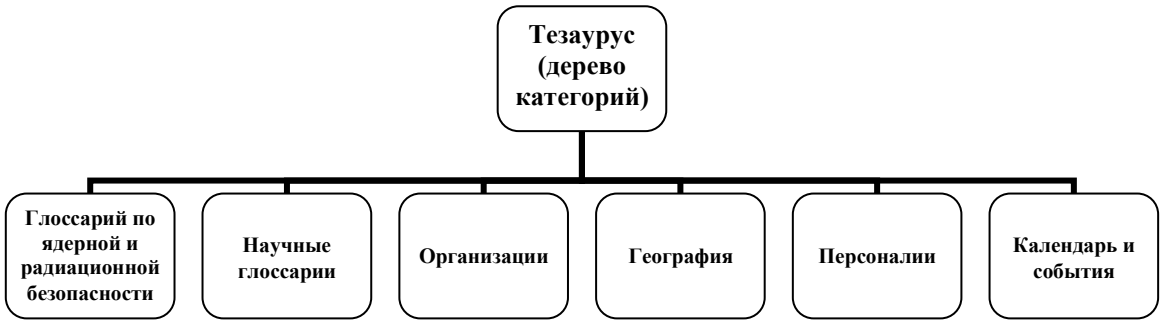

Рис. 1. Верхний уровень тезауруса

Fig. 1. Top level of the thesaurus

Глоссарий включает основные термины (русское и английское название) в области ядерной и радиационной безопасности с учетом белорусской специфики, а также менеджмента ядерных знаний, некоторых основ ядерной физики, физических единиц, основ администрирования и регулирования, учета ядерных материалов, работы с радиоактивными отходами и др. Белорусская специфика отражена через предложение значения терминов в национальных нормативно-правовых актах, а также расшифровку некоторых ключевых понятий и описание некоторых белорусских организаций, в том числе обладающих ядерными установками. Смысл предлагаемых терминов дается предельно кратко, только через главное определение. Дальнейшие подробности могут быть найдены по ссылкам на соответствующие термины.

Отличие глоссария по ядерной и радиационной безопасности от глоссария [16] заключается в преимущественном использовании терминологии белорусских национальных нормативно-правовых актов и выбранном одном значении каждого термина вне зависимости от языка.

В состав научных глоссариев BelNET входят глоссарии по физике, химии, информационным технологиям, техническим терминам, биологии, медицине, науках о Земле, общественным наукам (рис. 2). Изначально предполагалось, что объем создаваемых глоссариев не должен превышать 200 терминов, пригодных для использования в качестве ключевых слов.

В настоящее время специально для портала BelNET разработаны следующие глоссарии: «Единицы физических величин» (55 терминов), «Электричество и магнетизм» (129 терминов), «Квантовая физика» (85 терминов), «Атомная и ядерная физика» (200 терминов), «Астрофизика» (65 терминов), «Радиационное материаловедение, нанотехнологии» (124 термина), «Информационные технологии» (90 терминов), «Технические термины» (150 терминов). Некоторые термины встречаются в нескольких глоссариях, обеспечивая дополнительные горизонтальные отношения в тезаурусе.

Разделы «География», «Организации», «Персоналии», «Календарь и события» (рис. 1) помимо глоссариев включают справочники основных понятий, которые в составе соответствующих глоссариев играют роль предметных категорий.

В разделе «Организации» помимо всех необходимых справочников представлены следующие глоссарии: «Международные и межправительственные организации» (33 термина), белорусские – «Министерства и ведомства» (65 терминов), «Предприятия и организации» (26 терминов), «Научные организации» (30 терминов), «ВУЗы» (44 термина).

Раздел «География» включает подразделы «Регионы мира» и «Регионы Беларуси» (в том числе все административные районы по областям), а также важные географические объекты. Глоссарий «Страны мира» содержит записи о 176 странах мира – членах МАГАТЭ и КНДР, которая прекратила свое членство в МАГАТЭ в 1994 г.

Раздел «Персоналии» подразделяется на глоссарии «Зарубежные персоны» и «Белорусы». В эти глоссарии вошли данные о нобелевских лауреатах в области физики, химии и др. – специалистах в области ядерных знаний, ведущих ученых, руководителях ведомств, организаций и предприятий различного уровня.

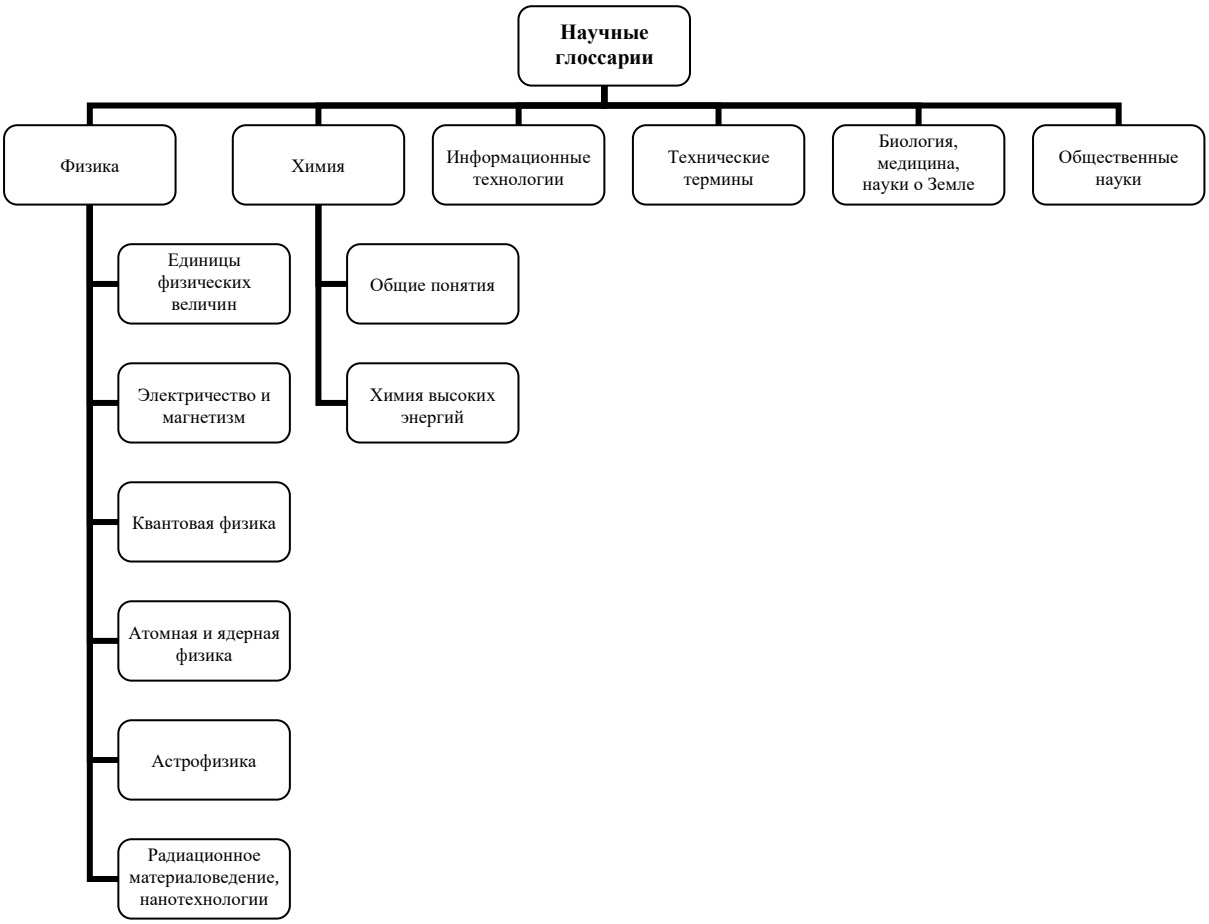

Рис. 2. Структура раздела «Научные глоссарии»
Fig. 2. Structure of section "Scientific glossaries"

Понятно, что разработка такого количества глоссариев является сложной задачей, и эта работа продолжается.

Методы. Алгоритм автоматического размещения ресурса в системе управления контентом eLab-Science. Главная идея алгоритма автоматического размещения ресурса в системе управления контентом eLab-Science заключается в использовании разработанного в системе управления контентом научно-образовательного портала eLab-Science [17, 18], на основе которой создан портал BelNET, полнотекстового поиска в отношении любых вновь создаваемых ресурсов, а также ресурсов, размещенных на портале ранее до внедрения семантических технологий. Это касается и записей в глоссариях.

В eLab-Science в кабинете создателя ресурса специальные кнопки «Индексировать» и «Систематизировать» позволяют пользователю на основе полнотекстового поиска по терминам всех глоссариев тезауруса получать автоматически предлагаемый системой список разделов портала, куда система рекомендует поместить материал, а также список ключевых слов.

На рис. 3 показана диаграмма декомпозиции eLab-Science в обозначениях IDEF1 [19] следующих компонентов системы и их связей: A1 – обслуживание пользователя, A2 – создание ресурса, A3 – автоматическая проверка ресурса, A4 – отображение ресурса на портале, A5 – систематизация ресурса (индексирование, систематизация ресурса, определение уровня доступа к ресурсу).

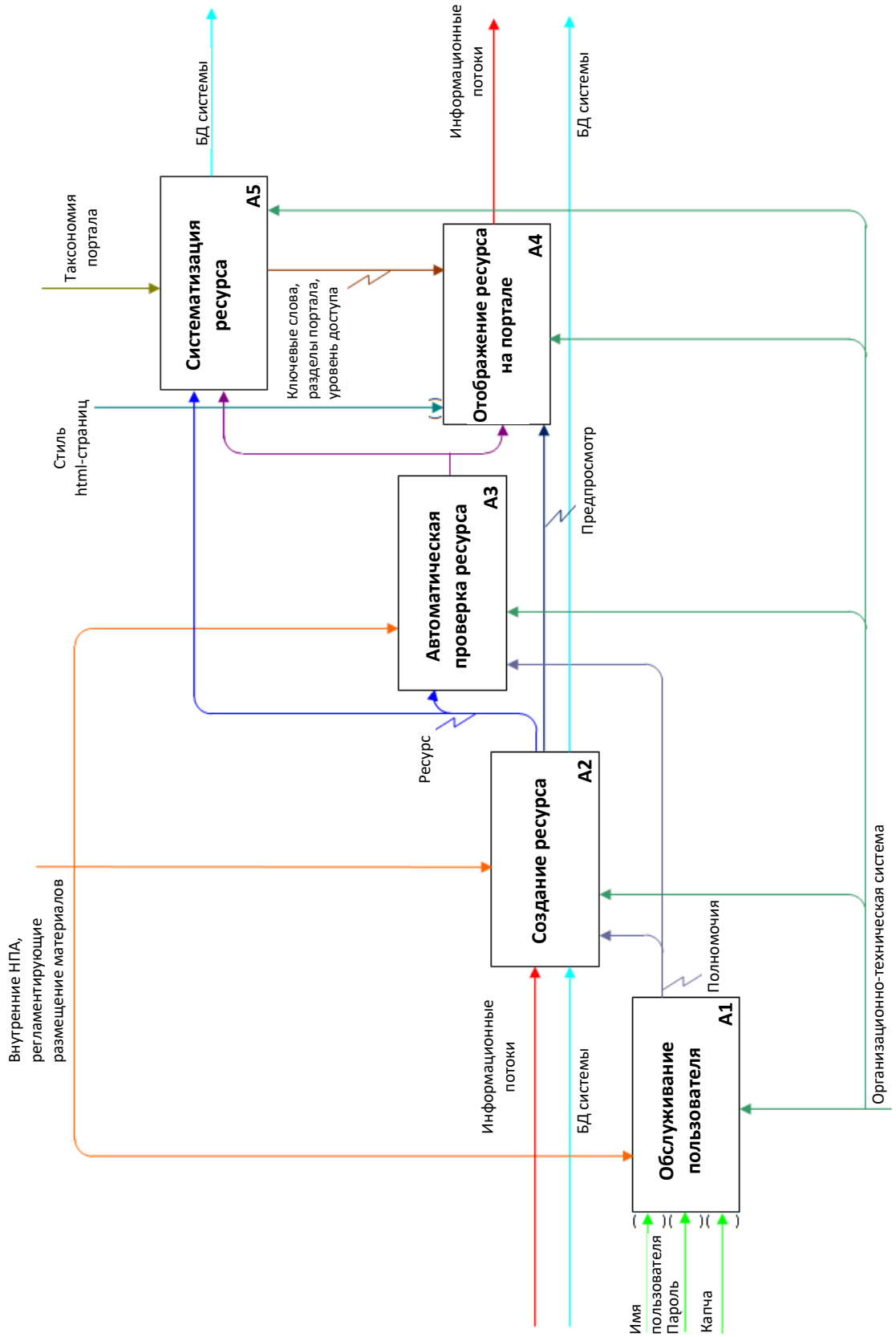

Рис. 3. Схема функциональной структуры портала BelNET в обозначениях IDEF1

Fig. 3. Diagram of the functional structure of portal BelNET in IDEF1 notation

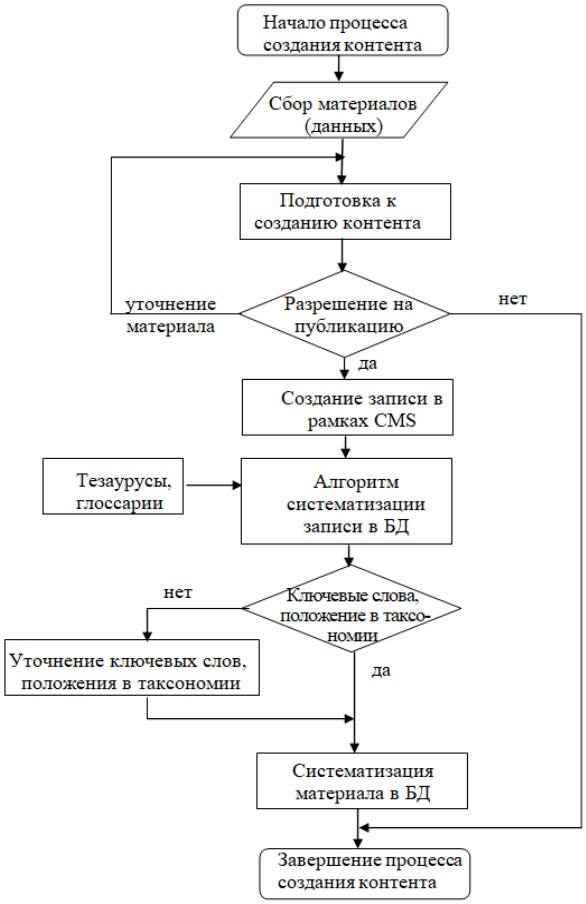

Рис. 4. Алгоритм размещения информационного ресурса на портале ядерных знаний
Fig. 4. Algorithm for posting an information resource on the nuclear knowledge portal

На рис. 4 изображен алгоритм в виде блок-схемы. Создатель новой записи (ресурса, материала, элемента глоссария) начинает ее размещение в своем кабинете, далее система производит автоматический анализ текста (полнотекстовый поиск по всем дескрипторам тезауруса) и пользователю предлагается утвердить набор ключевых слов и положение создаваемой записи в структуре портала (один или несколько разделов). Автор утверждает эти данные либо дополнительно предлагает свои варианты, после чего происходит окончательная систематизация ресурса на портале.

Результаты и обсуждение. Рассмотрим реализацию предложенного семантического алгоритма. Как было отмечено выше, глоссарий – это перечень терминов предметной области (ПрО) с их определениями. В состав глоссария, как правило, включаются термины, которые часто используются в узкой предметной области. Более широкий и системный перечень терминов и понятий ПрО называется тезаурусом. Глоссарий, как правило, составляется на основе ограниченного набора текстов ПрО и предназначен для решения некоторой частной задачи. Составление тезауруса – серьезная системная работа, требующая значительных ресурсов и привлечения экспертов ПрО.

В списке проблем, которые приходится решать при составлении тезауруса, можно указать следующие:

1. Полнота покрытия понятийного поля ПрО – необходимо охватить всю область знаний, не упустив ни одной части.

2. Верификация определений – сверка и согласование формулировок определений экспертами.

3. Структурирование терминов – организация понятийного поля в иерархическую, древовидную или другую систему взаимосвязей с учетом семантики ПрО.

4. Установка границ понятийного поля ПрО.

Отметим, что нередко разработчики тезауруса начинают ощущать себя энциклопедистами в стиле Дидро и д'Аламбера или пытаются превзойти Википедию. Необходимо учитывать, что полезный тезаурус – это терминология конкретной ПрО, собираемая и систематизируемая для решения конкретных практических задач. Попытка «объять необъятное» может ухудшить практическую полезность результатов работы и привести к неоправданным трудозатратам.

В разработанном и импортированном в информационную систему глоссарии (рис. 5) имеются следующие колонки:

– начальная буква – как в словаре;
– термин на русском языке;
– термин на английском языке;
– определения термина;
– ссылки на первоисточники;
– ссылки на категории рубрикатора.

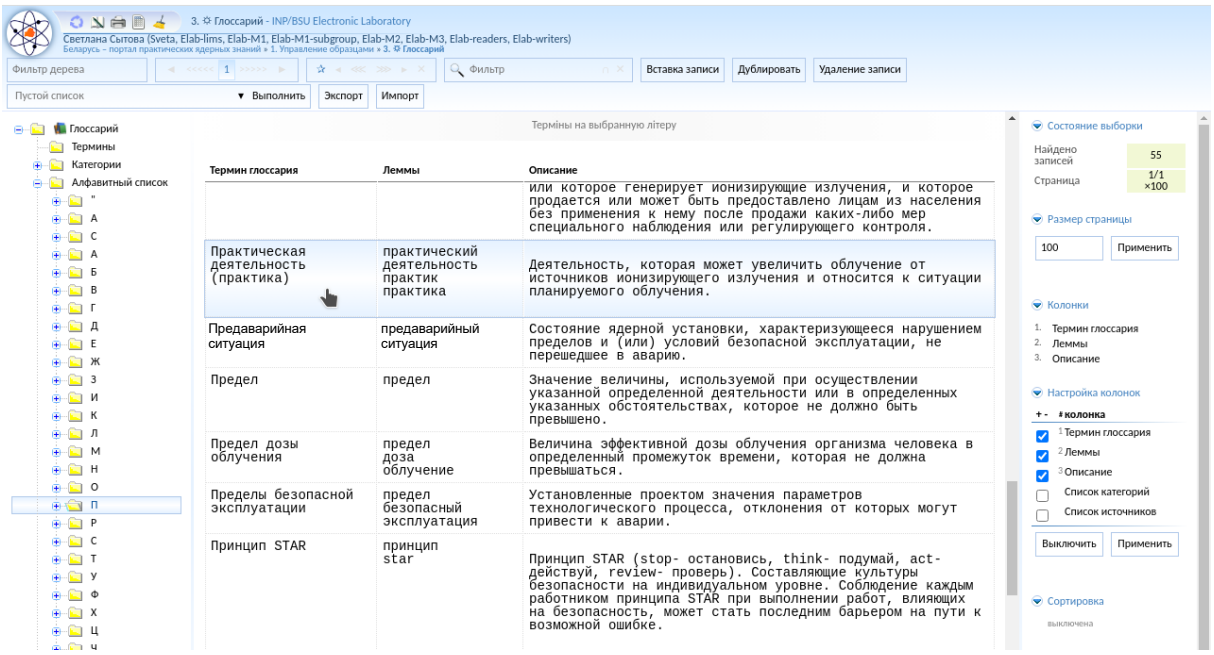

Рис. 5. Результаты импорта глоссария в информационную систему

Fig. 5. Results of the glossary import into the information system

Пример части списка категорий рубрикатора выглядит так:

1. Физические процессы и явления
   1.1. Общие понятия
   1.2. Единицы физических величин
2. Ионизирующие излучения
3. Атомное ядро и элементарные частицы
   3.1. Элементарные частицы
   3.2. Радионуклиды и химические элементы
   3.3. Ядерные процессы
4. Радиоактивность и радиоактивное вещество
5. Источники ионизирующего излучения (ИИИ)
   5.1. Закрытые ИИИ

5.2. Открытые ИИИ
5.3. Генерирующее оборудование
5.4. Работа с ИИИ.

На основе глоссария разработана концептуальная модель БД, в которой будет храниться вся информация для последующей автоматической обработки (рис. 6).

В приведенной модели требует пояснения только один момент – соотношение «Понятие» и «Синоним». Если посмотреть на рис. 5, то в колонке «Термин глоссария» таблицы формулировка термина имеет два варианта написания: практическая деятельность и практика. Такие случаи типичны для терминов рассматриваемой предметной области. В качестве примеров можно привести:

– тепловыделяющий элемент (ТВЭЛ);

– технико-экономическое обоснование (ТЭО);

– экспертиза безопасности в области использования атомной энергии и источников ионизирующего излучения (экспертиза безопасности).

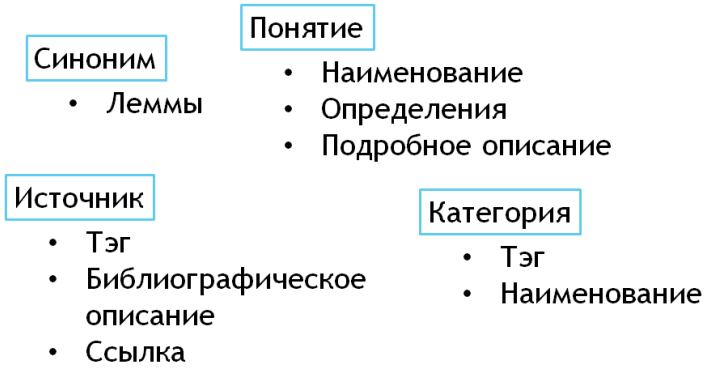

Рис. 6. Концептуальная модель БД глоссария
Fig. 6. Conceptual model of the glossary database

Таким образом, при поиске ключевых слов в тексте необходимо отдельно обнаруживать ТВЭЛ и «тепловыделяющий элемент», но соотносить их с одними и теми же определениями (понятиями) из ПрО. Соответственно, ТВЭЛ и «тепловыделяющий элемент» – это синонимы, которые концептуально в рамках модели относятся к понятию «тепловыделяющий элемент (ТВЭЛ)».

На рис. 6 особый интерес представляет колонка таблицы «Леммы». Это результат специальной обработки отдельных слов из состава понятия для последующего поиска данного понятия в тексте на естественном языке. Так, понятие «тепловыделяющий элемент (ТВЭЛ)» может упоминаться в тексте в различных словоформах в составе, например, следующих предложений:

– …. ТВЭЛы загружены в реактор …;

– …. в хранилище для тепловыделяющих элементов …

Все эти варианты использования понятия система должна находить в текстах и корректно обрабатывать. Для этого производится морфологический анализ слов понятия и выделяются леммы, т. е. проводится лемматизация.

Для поиска лемм, которые являются основой слова, его неизменяемой частью, выражающей лексическое значение, используются алгоритмы из состава библиотеки проекта PHPMorphy (https://github.com/cijic/phpmorphy). Словари этой библиотеки основаны на разработках проекта «Автоматическая обработка текста» (http://aot.ru/docs/sokirko/Dialog2004.htm). Если в словаре нет соответствующего слова, то применяется алгоритм стемминга – обработки словоформы с использованием эвристических правил для получения основы слова [20, 21]. В системе управления контентом eLab-Science применяется алгоритм Snowball (https://snowballstem.org/),

который поддерживает разные языки и реализован для большого числа средств разработки. Исходные тексты реализаций доступны в Интернете по открытым лицензиям.

В результате проблема словоформ решается, но возникает сложность с различными семантиками (смысловыми нагрузками) отдельных слов в тексте. В качестве примера можно привести анализ тестовой статьи «Атом» (рис. 7), где были выявлены следующие ключевые слова: конечное состояние, процесс, работа, система, УЕ, элементы.

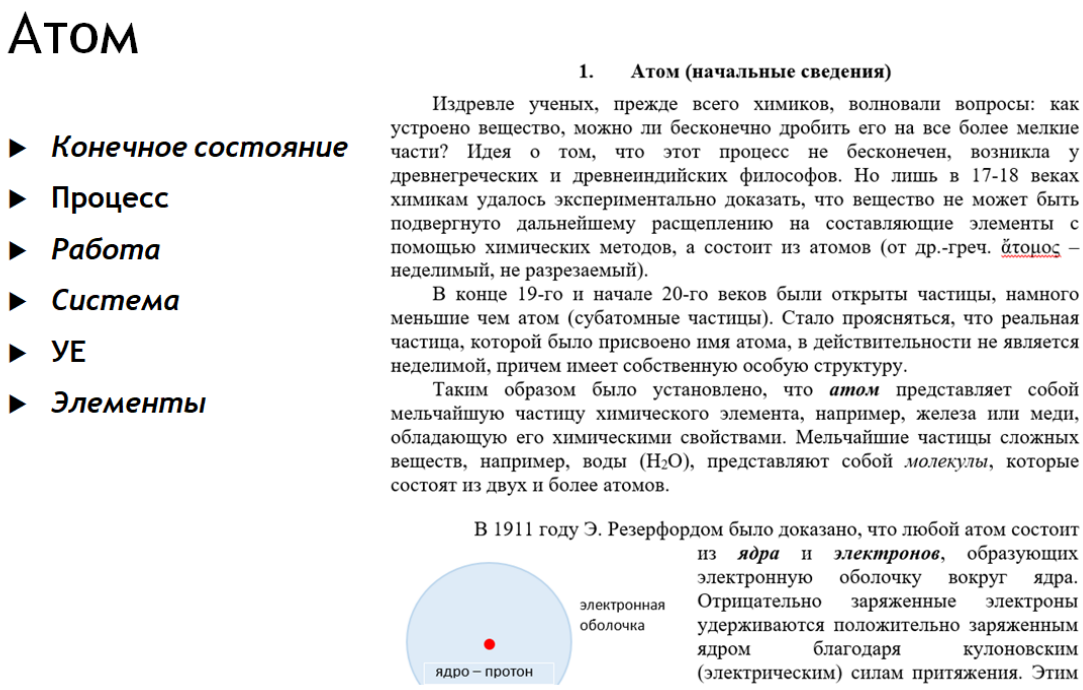

Рис. 7. Обработка тестовой статьи «Атом» с выделением ключевых слов

Fig. 7. Processing the test article "Atom" with the keywords highlighting

Со словом «процесс» получилась следующая семантическая коллизия. В глоссарии BelNET значение термина «процесс» выглядит так: «Процесс – последовательность действий или операций, в особенности ряд последовательных стадий изготовления продукта или некоторых других операций. Ряд взаимосвязанных или взаимодействующих операций, которые преобразуют вкладываемые ресурсы в конечные результаты».

В анализируемом тексте «Атом» есть фраза: «Издревле ученых, прежде всего химиков, волновали вопросы: как устроено вещество, можно ли бесконечно дробить его на все более мелкие части? Идея о том, что этот процесс не бесконечен, возникла у древнегреческих и древнеиндийских философов». То есть семантика в тексте и семантика в определении для слова «процесс» различаются.

С термином УЕ связана еще одна коллизия. В тексте статьи УЕ − это условная единица, а в глоссарии BelNET УЕ − это учетная единица (специальный термин из области обработки и хранения ядерных отходов).

На рис. 8 показан результат обработки тестовой статьи «Ядро». Здесь имеет место следующее нарушение семантики для термина «фон». В тексте статьи фраза с этим понятием звучит так: «…их взаимное электрическое отталкивание даже на фоне сильного ядерного притяжения». Слово «фон» тут является связующим внутри предложения («синтаксический сахар», как говорят лингвисты), но в рамках ПрО ядерных технологий понятие «радиационный фон» − это важнейший термин, который имеет строго определенную семантику и контекст применения.

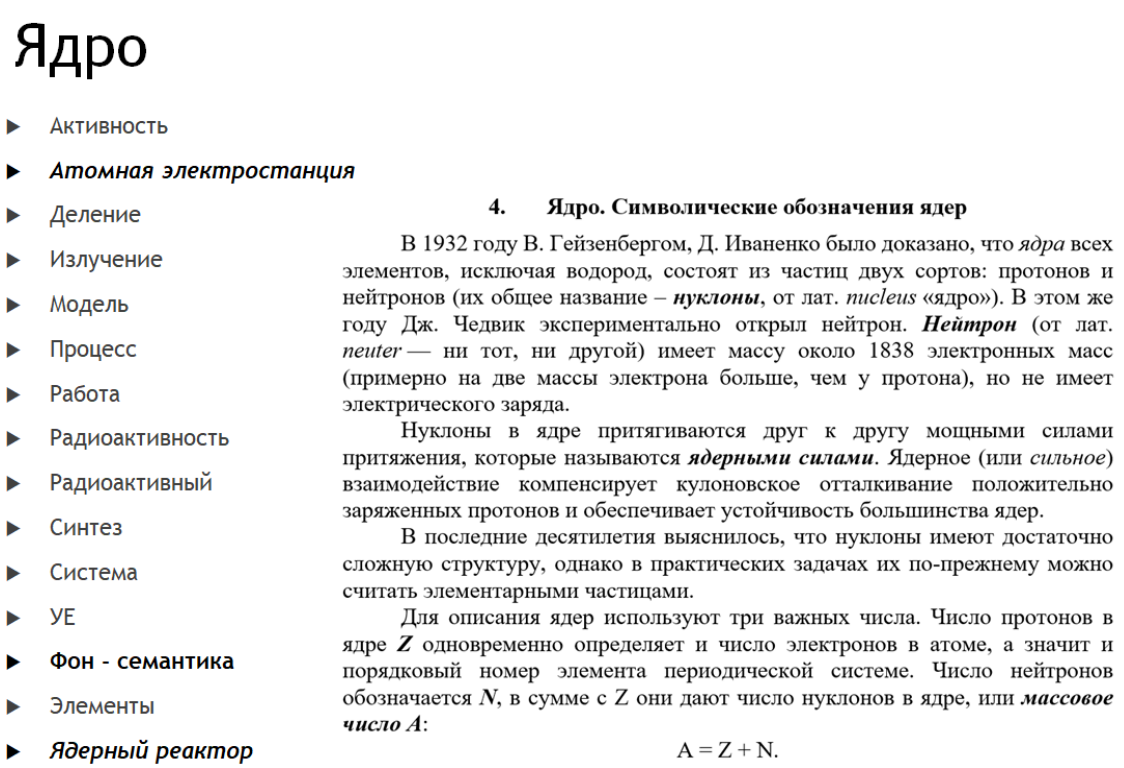

Рис. 8. Обработка статьи «Ядро» с выделением ключевых слов

Fig. 8. Processing the article "Nucleus" test with the keywords highlighting

Очевидно, что в вышеперечисленных и других аналогичных случаях проблему различия семантики в тексте и семантики термина из глоссария может решить только автор ресурса.

Вызывает интерес тот факт, что ключевые слова «атом» (рис. 7) и «ядро» (рис. 8) в данных текстах алгоритмом определены не были, так как тестируемый глоссарий по ядерной и радиационной безопасности попросту не содержит терминов «атом» и «ядро». Это означает, что для адекватной работы алгоритма необходимо использовать несколько тематических глоссариев.

Можно привести еще некоторое количество примеров подобного вида. На практике эта особенность преодолевается тем, что, как указывалось ранее, алгоритмы выделения ключевых слов работают в полуавтоматическом режиме. Список найденных ключевых слов предлагается пользователю, который сам выбирает, включать их в метаданные своего текста или не включать. Соответственно, пользователь из списка в левой колонке выберет то, что относится к его тексту, и исключит из предложенного списка те ключевые слова, которые не относятся к его предметной области.

Понятно, что глоссарии не должны содержать очень подробные «мелкие» термины. Это, с одной стороны, замедлит работу алгоритма из-за неоправданного объема глоссария, а с другой − выдаст большое количество терминов, которые не должны быть «ключевыми словами», и автоматически заставит пользователя исключать большое количество терминов из предложенного автоматически сформированного списка ключевых слов.

В отношении алгоритма определения положения создаваемого ресурса в таксономии портала начинает работать колонка «Список категорий», связанная с ключевым словом и с соответствующим одним или несколькими разделами в таксономии.

Оптимизация. Для проведения оптимизации алгоритма обработки задача, которая решается при поиске ключевых слов, может быть сформулирована как поиск пересечения нескольких цепочек лемм: цепочка лемм документа и набор цепочек лемм ключевых слов. В итоге необходима выборка цепочек, результаты пересечений которых имеют мощность (число элементов), отличную от нуля. В общем случае у этой задачи высокий уровень вычислительной сложности,

что может быть проблемой для больших текстов и ограничивать максимальный размер глоссария. В текущей реализации предлагаются следующие способы оптимизации: буферизация редко меняющихся данных и минимизация числа запросов к БД.

Буферизация редко меняющихся данных – это простой и широко используемый способ ускорения поисковых алгоритмов. Он основан на том, что после разработки содержимое глоссария и (или) тезауруса меняется очень редко. Такие изменения, как правило, представляют собой исправление отдельных ошибок, недоработок и опечаток. Оно проводится как процедура согласования между пользователем, который обнаружил недоработку, автором, внесшим соответствующий термин, и сотрудником, ответственным за сайт. Это в реальности не происходит динамически и за несколько секунд. Поэтому можно разрабатывать алгоритм обработки, предполагая, что в течение одного сеанса обработки данные глоссария меняться не будут и можно разместить их в буфере в оперативной памяти на старте алгоритма, предварительно подготовив данные для поиска вхождений цепочек лемм. Далее, пока обрабатывается текущая группа текстов, содержимое этого буфера не меняется.

Следует отметить, что объем данных типичного глоссария после подготовки невелик с точки зрения аппаратных возможностей современного компьютера: 12 Мб для глоссария из 525 терминов. Как аргумент для использования алгоритма обработки можно привести и то, что в случае обнаружения серьезной недоработки в каком-либо глоссарии либо внедрения нового глоссария потребуется обработка всех текстов заново. Это в любом случае приведет к перезапуску системы и полной перезагрузке всех данных.

Для минимизации числа запросов к БД было проведено испытание системы в разных аппаратных конфигурациях. Оно показало, что целесообразно загружать все данные, относящиеся к конкретному тексту сразу, т. е. лучше загрузить весь массив лемм документа, чем осуществлять фильтрацию по таблице отдельных ключевых слов.

При работе с реальным глоссарием формируются сотни (около 600 в конкретном тесте) простых запросов с фильтрацией по строковому значению. Эти запросы очень быстро выполняются сервером, но при этом каждый из них влечет за собой затраты на обмен данными: передачу запроса в БД и получение данных оттуда. Эта величина постоянна, но сотни повторений создают расход процессорного ресурса, который будет меньше, если применить альтернативный метод: загрузить все леммы текста и выполнить операции поиска в оперативной памяти, не привлекая БД.

В таблице и на рис. 9 показаны результаты нагрузочного испытания реализованного алгоритма для выяснения поведения системы на текстах различного объема. Тестируемый глоссарий содержит 525 терминов. В качестве проверочных текстов использовались реальные документы различных объемов и форматов из состава нормативной базы, которые размещены в открытом доступе на сайте Департамента по ядерной и радиационной безопасности Министерства чрезвычайных ситуаций Республики Беларусь (Госатомнадзор, https://gosatomnadzor.mchs.gov.by/), в интересах которого проводится данная работа.

Результаты испытаний работы системы под нагрузкой

Results of the system performance under load

Число символов *Number of symbols*	Число лемм в тексте *Number of lemmas in the text*	Число найденных терминов *Number of terms found*	Время обработки, с *Processing time, s*
15 438	1208	6	0,03
24 921	1951	15	0,05
182 764	17 669	43	0,45
4 870 803	165 478	113	45,50

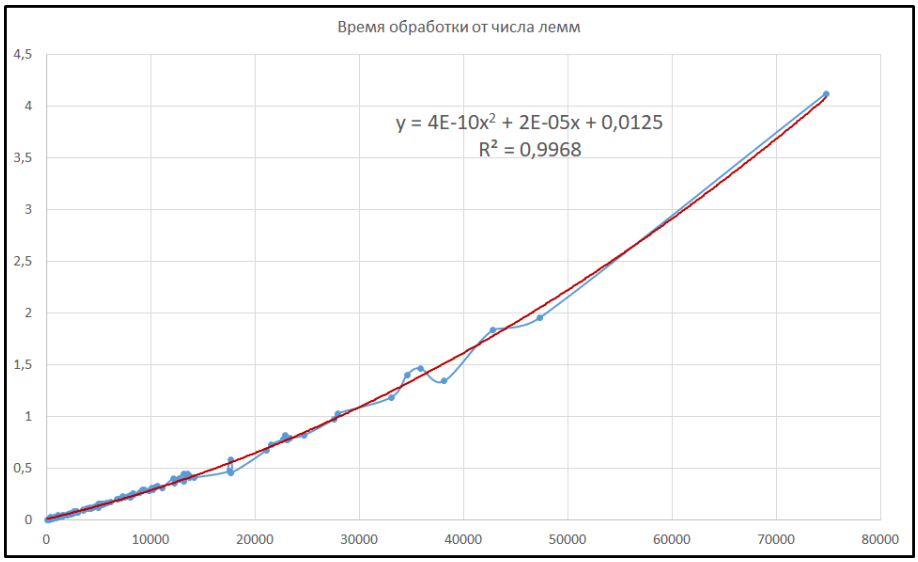

Рис. 9. Результаты испытаний работы системы и квадратичный тренд

Fig. 9. Results of the system performance and quadratic trend

На рис. 9 видно, что зависимость времени обработки от числа лемм скорее является квадратичной с $R^2 = 0{,}9968$, чем линейной с $R^2 = 0{,}9486$, при аппроксимации функцией $y = 4 \cdot 10^{-5}x - 0{,}08$.

Компьютер, на котором проводились испытания, был оснащен процессором Intel i5-7200U (тактовая частота 2,5 ГГц). Объем оперативной памяти практически не влиял на результат, так как требуемый объем ОЗУ не превысил 50 Мб. Оказалось, что даже для очень больших документов (в рассмотренном примере самый большой документ содержал около 200 страниц) время работы находится в приемлемых рамках – 4,1 с. Типичные документы – статьи и т. п. объемом несколько страниц – обрабатываются менее секунды.

Поскольку от рассмотренных алгоритмов не требуется мгновенный результат, работа может быть выполнена в фоновом режиме. Также очевидно, что размещение огромных документов одной записью зачастую не очень оптимально с точки зрения читабельности и наглядности материала. Понятно, что гораздо лучше использовать разбиение материала на главы и разделы.

Заключение. Портал BelNET постоянно развивается и бесперебойно функционирует с 2016 г. За прошедшее время число посетителей портала исчисляется тысячами, не только из Республики Беларусь и стран СНГ, но и со всего мира, со всех континентов, о чем свидетельствуют счетчики посещений портала, установленные в системе.

Подчеркнем, что процесс наполнения портала информацией и заполнения базы знаний, разработки специальных материалов для системы дистанционного обучения любого портала, тем более портала ядерных знаний, трудоемкий и длительный. И в этом смысле работа над BelNET находится в самом начале.

Можно констатировать также, что результаты реализации разработанных семантических алгоритмов, описанные в данной статье, являются очень хорошими, а их внедрение на портале BelNET позволяет отнести его к семантическим порталам.

Вклад авторов. *С. Н. Сытова* осуществила постановку проблемы, научное руководство ее решением, разработала структуру тезауруса и некоторые глоссарии, подготовила статью к публикации. *А. П. Дунец* реализовал семантические алгоритмы, включая полнотекстовый поиск в системе, а также провел оптимизацию работы и тестирование семантических алгоритмов. *А. Н. Коваленко* и *В. В. Гавриловец* разработали ядро системы eLab-Science. *С. В. Черепица* разработал структуру тезауруса и составил некоторые глоссарии.

Список использованных источников

1. Maintaining Knowledge, Training and Infrastructure for Research and Development in Nuclear Safety: INSAG-16. – Vienna : IAEA, 2003. – 19 p.

2. Knowledge Management for Nuclear Industry Operating Organizations. IAEA-TECDOC-1510. – Vienna : IAEA, 2006. – 185 p.

3. Knowledge Management and Its Implementation in Nuclear Organizations. IAEA Nuclear Energy Series No. NG-T-6. 10. – Vienna : IAEA, 2016. – 52 p.

4. Managing Nuclear Safety Knowledge: National Approaches and Experience. Safety Reports Series No. 105. – Vienna : IAEA, 2021. – 45 p.

5. Арсентьев, С. В. Корпоративная система современных ядерных знаний / С. В. Арсентьев // Глобальная ядерная безопасность. – 2023. – № 1(46). – С. 92–103.

6. Exploring Semantic Technologies and Their Application to Nuclear Knowledge Management. IAEA Nuclear Energy Series No. NG-T-6.15. – Vienna : IAEA, 2021. – 62 p.

7. Сытова, С. Н. Система управления ядерными знаниями в Республике Беларусь / С. Н. Сытова // Журнал БГУ. Физика. – 2022. − № 2. – С. 87–98.

8. Управление ядерными знаниями в системе научно-технической информации Республики Беларусь / С. Н. Сытова [и др.] // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2022) : докл. XXI Междунар. науч.-техн. конф., Минск, 17 нояб. 2022 г. – Минск : ОИПИ НАН Беларуси, 2022. – С. 265–269.

9. Nuclear knowledge management in the Republic of Belarus / S. Sytova [et al.] // Nonlinear Dynamics and Applications. – 2022. – Vol. 28. – P. 440–449.

10. Белорусский портал ядерных знаний BelNET: вчера, сегодня, завтра / С. Н. Сытова [и др.] // Сахаровские чтения 2023 года: экологические проблемы XXI века : материалы 23-й Междунар. науч. конф., Минск, Беларусь, 18–19 мая 2023 г. : в 2 ч. – Минск, 2023. – Ч. 2. – С. 158–162.

11. Онтологии и тезаурусы: модели, инструменты / Б. В. Добров [и др.]. − М. : Бином. Лаборатория знаний, 2009. − 173 с.

12. Лукашевич, Н. В. Тезаурусы в задачах информационного поиска / Н. В. Лукашевич. − М. : Изд-во МГУ, 2011. − 512 с.

13. Кириллович, А. В. Программная система для разработки многоязычного тезауруса / А. В. Кириллович, А. М. Баширов, А. Р. Гатиатуллин // Программные продукты и системы. – 2018. – Т. 31. – С. 112–120.

14. UNESCO SC/W/255. Guidelines for the Establishment and Development of Monolingual Thesauri. – N. Y. : UNESCO, 1973. − 37 p.

15. INIS Thesaurus. English Version. IAEA-INIS Reference Series. IAEA-INIS-01 (2018/09). – Vienna : IAEA, 2018. – 1312 p.

16. Михалевич, М. М. О лингвистических аспектах подготовки национального глоссария по ядерной и радиационной безопасности Республики Беларусь / М. М. Михалевич, Н. Н. Тушин // Сахаровские чтения 2022 года: экологические проблемы XXI века : материалы 22-й Междунар. науч. конф., Минск, Беларусь, 19–20 мая 2022 г. : в 2 ч. – Минск, 2022. – Ч. 1. – С. 176−179.

17. Информационная система eLab для аккредитованных испытательных лабораторий на основе свободного программного обеспечения / С. Н. Сытова [и др.] // Информатика. – 2017. − № 3(55). – С. 49–61.

18. Сытова, С. Н. Информационная система eLab в науке, практике, образовании / С. Н. Сытова. – Минск : Изд. центр БГУ, 2021. – 202 с.

19. Марка, Д. А. Методология структурного анализа и проектирования SADT : пер. с англ. / Д. А. Марка, К. МакГоуэн. – М. : Метатехнология, 1993. – 240 с.

20. Lovins, J. B. Development of a stemming algorithm / J. B. Lovins // Mechanical Translation and Computational Linguistics. – 1968. – Vol. 11. – P. 22–31.

21. Frakes, W. B. Strength and similarity of affix removal stemming algorithms / W. B. Frakes, C. J. Fox // SIGIR Forum. – 2003. – Vol. 37. – P. 26–30.

References

1. *Maintaining Knowledge, Training and Infrastructure for Research and Development in Nuclear Safety: INSAG-16*. Vienna, IAEA, 2003, 19 p.

2. *Knowledge Management for Nuclear Industry Operating Organizations*. *IAEA-TECDOC-1510*. Vienna, IAEA, 2006, 185 p.

3. *Knowledge Management and Its Implementation in Nuclear Organizations. IAEA Nuclear Energy Series No. NG-T-6. 10*. Vienna, IAEA, 2016, 52 p.

4. *Managing Nuclear Safety Knowledge: National Approaches and Experience. Safety Reports Series No. 105*. Vienna, IAEA, 2021, 45 p.

5. Arsentev S. V. *Corporate system of modern nuclear knowledge*. Global'naja jadernaja bezopasnost' [*Global Nuclear Security*], 2023, no. 1(46), pp. 92–103 (In Russ.).

6. *Exploring Semantic Technologies and Their Application to Nuclear Knowledge Management. IAEA Nuclear Energy Series No. NG-T-6.15*. Vienna, IAEA, 2021, 62 p.

7. Sytova S. N. *Nuclear knowledge management system in the Republic of Belarus*. Zhurnal Belorusskogo gosudarstvennogo universiteta. Fizika [*Journal of the Belarusian State University. Physics*], 2022, no. 2, pp. 87–98 (In Russ.).

8. Sytova S. N., Bartkevich A. R., Verenich K. A., Gavrilovec V. V., Gurachevskij V. L., ..., Cherepica S. V. *Nuclear knowledge management in the scientific and technical information system of the Republic of Belarus*. Razvitie informatizacii i gosudarstvennoj sistemy nauchno-tehnicheskoj informacii (RINTI-2022) : doklady XXI Mezhdunarodnoj nauchno-tehnicheskoj konferencii, Minsk, 17 nojabrja 2022 g. [*Development of Informatization and the State System of Scientific and Technical Information (RINTI-2022) : Reports of the XXI International Scientific and Technical Conference, Minsk, 17 November 2022*]. Minsk, Ob"edinennyj institut problem informatiki Nacional'noj akademii nauk Belarusi, 2022, pp. 265–269 (In Russ.).

9. Sytova S., Charapitsa S., Kavalenka A., Dunets A., Haurilavets V. Nuclear knowledge management in the Republic of Belarus. *Nonlinear Dynamics and Applications*, 2022, vol. 28, pp. 440–449.

10. Sytova S. N., Bartkevich A. R., Verenich K. A., Gavrilovec V. V., Dunec A. P., ..., Cherepica S. V. *Belarusian nuclear knowledge portal BelNET: yesterday, today, tomorrow*. Saharovskie chtenija 2023 goda: jekologicheskie problemy XXI veka : materialy 23-j Mezhdunarodnoj nauchnoj konferencii, Minsk, Belarus', 18–19 maja 2023 g. : v 2 chastjah [*Sakharov Readings 2023: Environmental Problems of the XXI Century : Materials of the 23rd International Scientific Conference, Minsk, Belarus, 18–19 May 2023 : in 2 Parts*]. Minsk, 2023, part 2, p. 158–162 (In Russ.).

11. Dobrov B. V., Ivanov V. V., Lukashevich N. V., Solov'ev V. D. Ontologii i tezaurusy: modeli, instrumenty. *Ontologies and Thesauri: Models, Tools.* Moscow, Binom. Laboratoriya znanij, 2009, 173 p. (In Russ.).

12. Lukashevich N. V. Tezaurusy v zadachah informacionnogo poiska. *Thesauruses in Information Retrieval Tasks.* Moscow, Izdatel'stvo Moskovskogo gosudarstvennogo universiteta, 2011, 512 p. (In Russ.).

13. Kirillovich A. V., Bashirov A. M., Gatiatullin A. R. *Software system for developing a multilingual thesaurus*. Programmnye produkty i sistemy [*Software & Systems*], 2018, vol. 31, pp.112–120 (In Russ.).

14. *UNESCO SC/W/255. Guidelines for the Establishment and Development of Monolingual Thesauri*. New York, UNESCO, 1973, 37 p.

15. *INIS Thesaurus. English Version. IAEA-INIS Reference Series. IAEA-INIS-01 (2018/09).* Vienna, IAEA, 2018, 1312 p.

16. Mihalevich M. M., Tushin N. N. *On the linguistic aspects of the preparation of the national glossary on nuclear and radiation safety of the Republic of Belarus*. Saharovskie chtenija 2022 goda: jekologicheskie problemy XXI veka : materialy 22-j Mezhdunarodnoj nauchnoj konferencii, Minsk, Belarus', 19–20 maja 2022 g. : v 2 chastjah [*Sakharov Readings 2022: Environmental Problems of the XXI Century : Materials of the 22nd International Scientific Conference, Minsk, Belarus, 19–20 May 2022 : in 2 Parts*]. Minsk, 2022, part 1, pp. 176−179 (In Russ.).

17. Sytova S. N., Dunets A. P., Kovalenko A. N., Mazanik A. L., Sidorovich T. P., Charapitsa S. V. *Information system eLab for accredited testing laboratories*. Informatika [*Informatics*], 2017, no. 3(55), pp. 49–61 (In Russ.).

18. Sytova S. N. Informacionnaya sistema eLab v nauke, praktike, obrazovanii. *Information System eLab in Science, Practice, Education.* Minsk, Izdatel'skij centr Belorusskogo gosudarstvennogo universiteta, 2021, 202 p. (In Russ.).

19. Marca D. A., McGowan C. L. *Sadt: Structured Analysis and Design Techniques*. McGraw-Hill, 1987, 392 p.

20. Lovins J. B. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 1968, vol. 11, pp. 22–31.

21. Frakes W. B., Fox C. J. Strength and similarity of affix removal stemming algorithms. *SIGIR Forum*. 2003, vol. 37, pp. 26–30.

Информация об авторах

Сытова Светлана Николаевна, кандидат физико-математических наук, доцент, заведующий лабораторией, Институт ядерных проблем Белорусского государственного университета.
E-mail: sytova@inp.bsu.by
https://orcid.org/0000-0002-2476-9979

Гавриловец Виктор Васильевич, научный сотрудник, Институт ядерных проблем Белорусского государственного университета.
E-mail: bycel@tut.by
https://orcid.org/0000-0002-9452-7465

Дунец Андрей Петрович, старший научный сотрудник, Институт ядерных проблем Белорусского государственного университета.
E-mail: dunets@gmail.com
https://orcid.org/0009-0006-0980-7697

Коваленко Антон Николаевич, старший научный сотрудник, Институт ядерных проблем Белорусского государственного университета.
E-mail: anton.kavalenka@gmail.com
https://orcid.org/0000-0002-0320-2092

Черепица Сергей Вячеславович, кандидат физико-математических наук, доцент, ведущий научный сотрудник, Институт ядерных проблем Белорусского государственного университета.
E-mail: svcharapitsa@gmail.com
https://orcid.org/0000-0001-9657-1948

Information about the authors

Svetlana N. Sytova, Ph. D. (Phys.-Math.), Assoc. Prof., Head of the Laboratory, Institute for Nuclear Problems of the Belarusian State University.
E-mail: sytova@inp.bsu.by
https://orcid.org/0000-0002-2476-9979

Viktar V. Haurylavets, Researcher, Institute for Nuclear Problems of the Belarusian State University.
E-mail: bycel@tut.by
https://orcid.org/0000-0002-9452-7465

Andrei P. Dunets, Senior Researcher, Institute for Nuclear Problems of the Belarusian State University.
E-mail: dunets@gmail.com
https://orcid.org/0009-0006-0980-7697

Anton N. Kavalenka, Senior Researcher, Institute for Nuclear Problems of the Belarusian State University.
E-mail: anton.kavalenka@gmail.com
https://orcid.org/0000-0002-0320-2092

Siarhei V. Charapitsa, Ph. D. (Phys.-Math.), Assoc. Prof., Leading Researcher, Institute for Nuclear Problems of the Belarusian State University.
E-mail: svcharapitsa@gmail.com
https://orcid.org/0000-0001-9657-1948