УДК 004.4; 004.6; 004.9

ИСПОЛЬЗОВАНИЕ СЕМАНТИЧЕСКИХ ТЕХНОЛОГИЙ ДЛЯ РАЗВИТИЯ ПОРТАЛА ЯДЕРНЫХ ЗНАНИЙ BELNET

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

Институт ядерных проблем Белгосуниверситета,
ул. Бобруйская 11, 220006, г. Минск, Беларусь, sytova@inp.bsu.by

Рассмотрены теоретические основы семантических технологий, а также алгоритмы, предложенные для дальнейшего развития белорусского портала ядерных знаний BelNET https://belnet.bsu.by/.

Ключевые слова: семантические технологии; таксономия; ядерные знания; система управления знаниями; информационная система.

USING SEMANTIC TECHNOLOGIES TO DEVELOP THE PORTAL OF NUCLEAR KNOWLEDGE BELNET

S.N. Sytova, V.V. Haurilavets, A.P. Dunets, A.N. Kavalenka, S.V. Charapitsa

Institute for Nuclear Problems, Belarusian State University,
st. Bobruiskaya 11, 220006, Minsk, Belarus, sytova@inp.bsu.by

The theoretical foundations of semantic technologies, as well as the algorithms proposed for the further development of the Belarusian portal of nuclear knowledge BelNET https://belnet.bsu.by/, are considered.

Keywords: semantic technologies; taxonomy; nuclear knowledge; knowledge management system; information system.

Введение

Международное агентство по атомной энергии (МАГАТЭ) уделяет пристальное внимание управлению (менеджменту) в области ядерных знаний [1, 2], имеющего большое значение для развития и сохранения необходимых технических знаний, опыта и успешной реализации различных ядерно-физических, ядерно-энергетических программ и современных ядерных технологий. Одним из наиболее действенных инструментов в менеджменте ядерных знаний являются порталы ядерных знаний.

В Беларуси в настоящее время формируется полноценная система управления ядерными знаниями (СУЯЗ), основу которой составляет

портал ядерных знаний BelNET (Belarusian Nuclear Education and Training Portal, https://belnet.bsu.by/) [3, 4]. Его цели полностью соответствуют подходам МАГАТЭ к менеджменту ядерных знаний.

Данная статья посвящена использованию семантических технологий на портале BelNET.

1. Теоретические основы семантических технологий

Использование семантических технологий при создании СУЯЗ позволяет получить целостный системный взгляд на предметную область, подойти с единых позиций к размещению разнородных материалов, а также с научной точки зрения – восстановить логические цепочки и связи между понятиями в предметной области [5, 6].

Назовем основные понятия семантических технологий, которые используются в работе: онтология, глоссарий, тезаурус, таксономия.

Онтология используется для подробной формализации области знаний с помощью концептуальной схемы, состоящей из структуры данных, содержащей все релевантные классы объектов, их связей, а также правил (теоремы, ограничения), принятых в этой области. Среди главных сфер применения онтологий следует назвать моделирование бизнес-процессов, искусственный интеллект и семантическую паутину. Семантическая паутина (https://www.w3.org/standards/semanticweb/) – это своеобразная надстройка над Всемирной паутиной (интернетом), облегчающая машинную обработку интернет-информации.

Данные, онтология и правила вывода, организованные в единую систему, образуют базу знаний [7] 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.

Тезаурусом называется словарь с дополнительными отношениями, охватывающий понятия, определения и термины области знаний или сферы деятельности, подчиняющиеся семантическим отношениям между терминами (синонимы, антонимы, паронимы, гипонимы, гиперонимы и т. п.). Обычные простейшие таксономические отношения в тезаурусах составляют несколько уровней отношений типа выше–ниже [8]. Специализированный тезаурус может разрабатываться как экспертами, так и построен с помощью специальных программных средств [9]. Для создания тезаурусов существует специальные ГОСТ [10] и ИСО ([11] и др.). На основе тезауруса может быть создана таксономия (иерархическая

структура) портала. Отметим правило [12], что если в тезаурусе нет подходящего дескриптора для поиска полезного понятия, следует предложить и ввести в тезаурус новый.

Для онлайн-поиска в поисковых запросах могут использоваться контролируемые термины (или дескрипторы) – это ключевые слова, произвольные текстовые слова или их комбинации, 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.

Для выбора релевантных ссылок из результатов поиска очень полезными элементами являются реферат, заголовок и дескрипторы (ключевые слова) исследуемого ресурса.

В применении к созданию информационного портала таксономия – иерархическая структура портала, которая может быть построена на основании семантических технологий, в частности, на основании одного или нескольких тезаурусов [13].

2. Подход МАГАТЭ и портала «Атомная энергия 2.0» к использованию семантических технологий

Согласно материалам МАГАТЭ [13], семантическая технология, лежащая в основе веб-поиска и управления онлайн-информацией, может и должна использоваться в ядерной области, чтобы помочь экспертам и заинтересованным сторонам поддерживать, сохранять и обмениваться ядерными знаниями. В настоящее время МАГАТЭ изучает различные прототипы и инициативы в области семантических технологий, которые могут принести пользу в области менеджмента ядерных знаний.

Цель больших информационных систем, таких как созданные под эгидой МАГАТЭ INIS https://www.iaea.org/resources/databases/inis (The International Nuclear Information System – Международная ядерная информационная система) или ETDE (Energy Technology Data Exchange – Обмен данными по энергетическим технологиям), состоит в том, чтобы облегчить поиск информации эффективным и экономичным способом. При предметной классификации (категоризации) каждая запись в базах данных (например, в INIS и ETDE) должна быть отнесена к определенной предметной категории. Кроме того, в зависимости от тематического содержания создаваемой записи может потребоваться присвоение одной или нескольких вторичных тематических категорий.

Российский портал «Атомная энергия 2.0» под эгидой госкорпорации Росатом https://www.atomic-energy.ru/ позиционирует себя как научное

семантическое средство массовой информации в области ядерных знаний, в которой используется семантическая библиотека из 2500 ключевых слов. В недавней презентации создателей портала «Атомная энергия 2.0» (https://www.atomic-energy.ru/news/2021/11/16/119447) есть данные о более 120 000 публикаций портала, каждая из которых была *вручную семантически отсортирована* по общей тематике (1000+ терминов), ключевым словам (1500+ терминов), географии (1000+ терминов), организациям (2000+ наименований), персоналиям (1500+ наименований) и событиям (1000+ наименований).

Однако ручная семантическая сортировка такого количества публикаций вызывает огромные вопросы как к объему человеко-часов, затраченных на данную работу, так и к системе управления контентом портала, не предоставляющей инструментов для автоматизации этого процесса.

3. Алгоритм автоматического размещения ресурса в BelNET

При работе над таксономией портала BelNET было принято решение придерживаться комбинированного подхода с использованием наработок МАГАТЭ и большого багажа знаний белорусских экспертов. Таксономия портала создается на основе нескольких верхних уровней специально разработанного тезауруса. На начальном этапе используется экспертное формирование тезауруса, а дальше используется алгоритм, описанный ниже. Тезаурус включает несколько глоссариев по различным темам, в том числе в области теоретической и ядерной физики, технический глоссарий, глоссарий по информационным технологиям, по общепринятой терминологии и др. Объем этих глоссариев не должен быть большим. При необходимости внесения нового термина глоссарии всегда могут быть дополнены. Оптимальный объем тезауруса в настоящий момент представляется равным около 2000 терминов.

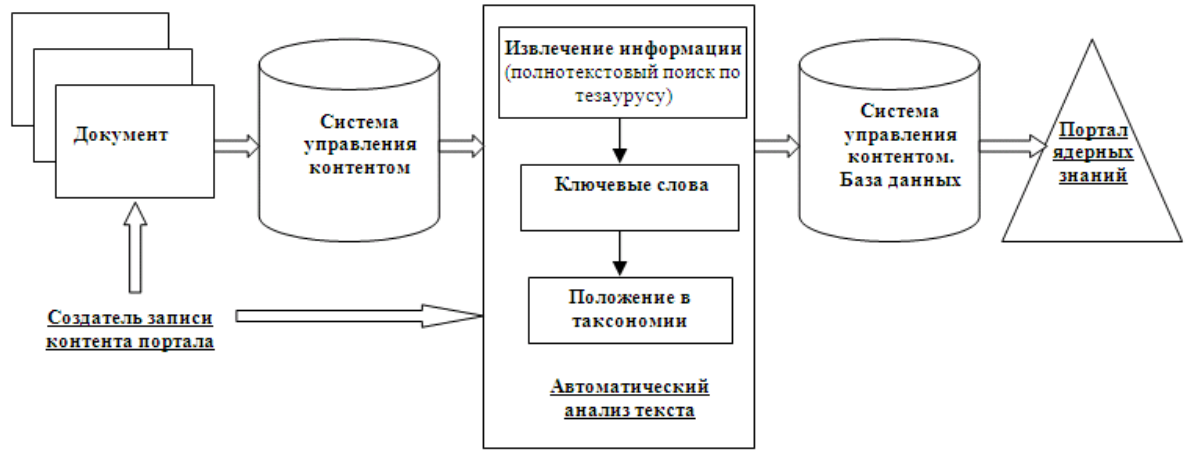
Главная идея использовать разработанный в системе управления контентом eLab-Science [4], на основе которой создан портал BelNET, полнотекстовый поиск в отношении любых вновь создаваемых ресурсов, а также ресурсов, размещенных на портале ранее до внедрения семантических технологий. Это касается и записей в глоссариях.

К такому решению привел отрицательный опыт при систематизации вручную ресурсов на портале BelNET, в результате чего многие разделы таксономии (а она была разработана в 2014 г. очень большой) до сих пор остаются пустыми, хотя, очевидно, многие ресурсы могли быть размещены в них и стать более доступными для читателей.

В eLab-Science в кабинете создателя ресурса специальные кнопки «Индексировать», «Систематизировать» позволяют пользователю получить ав-

томатически предлагаемый системой список разделов портала, куда система предлагает поместить материал, а также список ключевых слов.

Рисунок демонстрирует алгоритм, когда 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.



Алгоритм размещения информационного ресурса на портале ядерных знаний

Заключение

В настоящий момент идет активная работа над оригинальным тезаурусом, который должен содержать в обязательном порядке термины из белорусской специфики в области ядерных знаний.

В результате развития и внедрения семантических технологий на белорусском портале BelNET создается база знаний в области ядерных знаний – база ядерных знаний.

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 годы.

Библиографические ссылки

1. Managing Nuclear Safety Knowledge: National Approaches and Experience. Safety Reports Series № 105 STI/PUB/1938 ¦ 978-92-0-104221-7. Vienna: IAEA, 2021. 45 p.
2. Fast reactor knowledge preservation system: taxonomy and basic requirements. Vienna: International Atomic Energy Agency, 2008. 79 p.
3. Сытова С.Н. Система управления ядерными знаниями в Республике Беларусь. Журнал БГУ. Физика. 2022. № 2. С. 87–98. DOI: 10.33581/2520-2243-2022-2-87-98.
4. Сытова С.Н., Дунец А.П., Коваленко А.Н., Мазаник А.Л., Сидорович Т.П., Черепица С.В. Информационная система eLab для аккредитованных испытательных лабораторий на основе свободного программного обеспечения // Информатика. 2017. № 3. С. 49–61.
5. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М.: Изд-во МГУ, 2011. 512 с.
6. Добров Б.В. Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: модели, инструменты. М. : Бином. Лаборатория знаний, 2009. 173 с.
7. Sowa J.F. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Boston: Brooks/Cole, 2000. 594 p.
8. Загорулько Ю.А., Загорулько Г.Б. Онтологический подхлод к созданию научных интернет-ресурсов // Труды межд. конф. OSTIS-2015. С. 177–182.
9. Кириллович А.В., Баширов А.М., Гатиатуллин А.Р. Программная система для разработки многоязычного тезауруса. Программные продукты и системы // Software & Systems. 2018. № 31(1). С. 112–120.
10. ГОСТ 7.25-2001. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления. М.: Издательство стандартов. 2001. 14 с.
11. ИСО 25964-1:2011. Информация и документация. Тезаурусы и взаимосвязь с другими словарями. Часть 1. Тезаурусы для выдачи. М., 2011. 160 с.
12. UNESCO SC/W/255. Guidelines for the Establishment and Development of Monolingual Thesauri. M., 1973. 16 p.
13. IAEA Nuclear Energy Series NG-T-6.15. Exploring Semantic Technologies and Their Application to Nuclear Knowledge Management. Vienna: IAEA, 2021. 62 p.