Лингвистическое обеспечение портала НТИ Госкорпорации «Росатом»

Linguistic support of sci-tech information portal of ROSATOM National Corporation

Н. В. Игнатова, Е. Н. Ксионда, Т. Ф. Марова
Федеральное государственное унитарное предприятие
«Российский федеральный ядерный центр –
Всероссийский научно-исследовательский
институт экспериментальной физики»,
Саров, Россия

Natalya Ignatova, Elena Ksionda and Tatiana Marova
Russian Federal Nuclear Center –
All-Russian Research Institute of Experimental Physics,
Sarov, Russia

Портал НТИ – корпоративная электронная библиотека, создаваемая для оперативного обеспечения несекретной научно-технической информацией специалистов Госкорпорации путём предоставления удаленного доступа сотрудникам организаций. Важнейшей задачей библиотеки НТИ является объединение и сохранность всей отраслевой информации. Инструментом, позволяющим облегчить поиск и раскрыть содержание хранящихся источников информации, являются лингвистические средства.
Ключевые слова: лингвистические средства, тезаурус, портал НТИ Госкорпорации «Росатом».

The sci-tech information portal is the corporate digital library being built for efficient provision of unclassified information and remote access to this information for the National Corporation staff. The Sci-tech Library's main task is to acquire and preserve the whole of the branch information. Linguistic tools are to facilitate information retrieval and to reveal the content of preserved information sources.
Keywords: linguistic instruments, thesaurus, ROSATOM National Corporation sci-tech information portal.

В соответствии с концепцией и целевой моделью Системы Управления Знаний Госкорпорации «Росатом» портал НТИ выбран в качестве инструмента сохранения отраслевых знаний. Накопленный к настоящему времени материал портала НТИ, в совокупности с планируемым ростом его объема за счет пополнения публикаций и изданий организаций Госкорпорации «Росатом» определяет необходимость лингвистического обеспечения портала. Для взаимного обмена научно-исследовательской информацией между специалистами разных областей знаний необходимо нормализованное использование терминологии, с учётом основных Законов, Законодательных и нормативных актов, ГОСТов. Роль такого нормализованного терминологического собрания должен выполнить тезаурус. Тезаурус один из основных инструментов систематизации материалов портала НТИ, как средство поиска научно-технической информации.

Что такое тезаурус? Приводятся различные определения информационно-поисковых тезаурусов в различных источниках. Тезаурус – это словарь, предназначенный для поиска слов какого-либо языка по их смыслу. Тезаурусом могут называть список, классификацию.

Изучив свои фонды и фонды библиотек связанных по межбиблиотечному абонементу, выяснилось, что книг по созданию и ведения тезауруса невелико, и за информацией необходимо обращаться к периодическим изданиям, и к сайтам библиотек в Интернет.

Тезаурусов в фонде библиотеки всего три, и они различны по способу создания и исполнения, как различны и определения тезауруса. Это может быть:

- перечень терминов узконаправленной отрасли; Тезаурус информационно поисковый по приборостроению, средствам автоматизации и системам управления/ Отв. ред Изюмская К.М., Синицина Н.И. ЦНИИиТЭИ приборостроения и НИИ УМС г. Пермь – М. -1974.-251с.
- словарь терминов (Рождественский,Ю.В. Словарь терминов (Общеобразовательный тезаурус) [Текст]: общество. Семиотика. Экономика. Культура. Образование. / Ю. В. Рождественский. – М.: Флинта, 2002. – 112 с.;

- тезаурус информационно поисковый (Тезаурус по атомной науке и технике: лексико-семантический указатель. М.1983. Ч.1,2 (Госуд. комитет по использованию атомной энергии СССР. ЦНИИТЭИатоминформ).

Изучив всю доступную на тот момент информацию, опираясь на ГОСТ 7.25-2001 «Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления» можно приступать к работе по созданию тезауруса. Для этого необходимо проанализировать массив документов; определить тематический охват тезауруса; собрать массив лексических единиц; выбрать программное обеспечение и приступить к построению тезауруса.

Анализ массива документов

В рамках работы над проектом был проанализирован архив документированной информации портала НТИ. Архив анализировался на предмет выявления наиболее информативных материалов, отражающих результаты интеллектуальной деятельности корпорации. Состав предоставленных документов можно разделить на несколько составляющих: статьи сотрудников для научно-технических журналов (37%), доклады на общероссийские и международные конференции (26%), объекты интеллектуальной собственности (12%), отчеты об исследованиях и разработках (9%), учебно-методические материалы (4,2%), сведения о характеристиках материалов и оборудования (3,6%), монографии и труды, издаваемые предприятиями (3,2%), результаты экспериментов(3,9%), периодические журналы (0,8%), препринты (0,3%).

Определение тематического охвата Тезауруса

Все отобранные материалы, распределённые по таблицам, были проиндексированы и систематизированы по основным правилам и методам предложенным ГОСТом 7.59-2003. Индексирование документов. Общие требования к систематизации и предметизации. Индексация проходила по Универсальной десятичной классификации (УДК), принятой в научных и научно-технических библиотеках страны на основе выработанной методики. В процессе работы методические правила уточнялись и детализировались. Рубрицирование осуществлялось по Государственному рубрикатору научно-технической информации (ГРНТИ) на основании содержания документа. Предметная рубрика отражала максимально полно и точно содержание документа. Количество предметных рубрик ограничивалось не более 3.После анализа массива были выявлены следующие ошибки в индексации:

- Пропуски и перестановки символов (шифр 621.039.5 мог быть 621.0395 или 621.39.5 или 621.093.5);
- Пропуск одного из индексов;
- Замена индекса УДК на ГРНТИ или наоборот;
- Разница в трактовке текстов документов разными библиотекарями – тексты связанные одной тематикой одного автора индексировались по-разному;
- Пропуск индексов вспомогательных таблиц (общие определители формы документа – патента, таблицы, каталоги и т.д.)

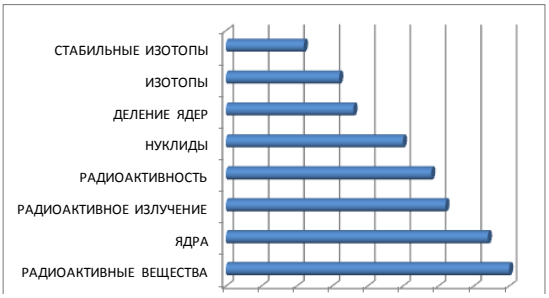
Классифицирование документов осуществлялось по специальному отраслевому классификатору НТИ. Классификатор содержит более 1500 рубрик, отражающих основные направлениям деятельности отрасли, состоит из четырех разделов Технологии атомной энергетики, Технологии неатомной энергетики, Технологии для неэнергетических рынков, Технологии неядерных вооружений.

По результатам проделанной работы были выявлены предметные рубрики для определения тематического охвата тезауруса портала НТИ. Первоначально это было семь предметных рубрик: атомная энергетика и промышленность, физика, химия, химические технологии, лазерная техника, вычислительная техника, ядерная техника которые соответствовали тематическому содержанию портала НТИ. Затем в рамках планируемого роста его объёма за счет пополнения организациями Госкорпорации «Росатом» и с опорой на разрабатываемый отраслевой классификатор НТИ количество областей было увеличено, добавлены области материаловедение, безопасность и аварии, ядерная медицина. Каждой предметной рубрика соответствовал свой массив документов, обработанный в табличной форме с УДК, ГРНТИ, ключевыми словами. Рис.1

Сбор массива лексических единиц

Следующий этап работы анализ ключевых слов – для каждой предметной рубрики рассматривалась однозначность термина, распространённость, краткость. Анализ выявил следующие недочёты допущенные библиотекарями при написание ключевых слов: замена букв кириллицы на латиницу, перестановка букв, пропуск букв, пробелы до термина и после термина, слова или словосочетания приводились и в единственном и во множественном числе.

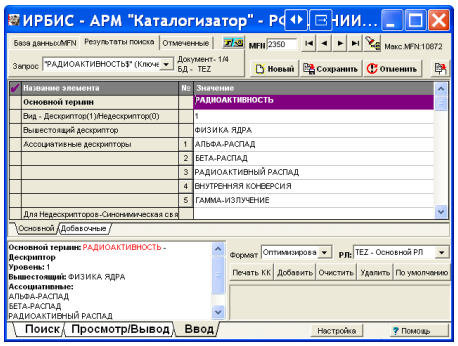
После исправления недочётов, в результате работы несложной программы, была выявлена частота встречаемости в текстах документов ключевых слов рубрики. Результата работы программы иллюстрирован см. рис.2



Не все ключевые слова стали терминами тезауруса. Были отобраны ключевые слова, отражающие смысловое содержание документа.

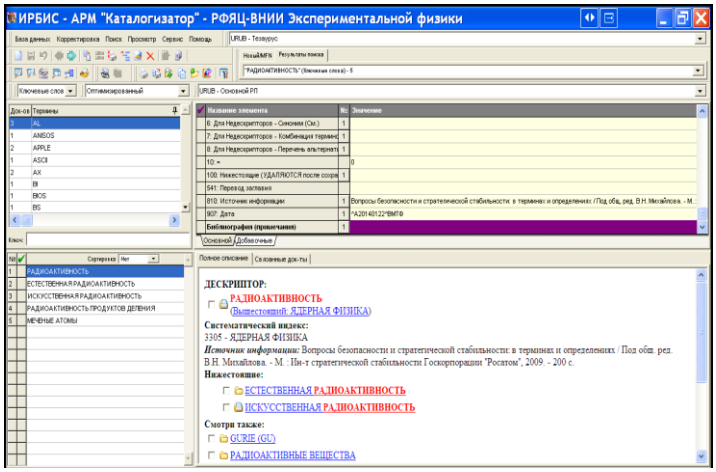
Выбор программного обеспечения

Библиотека РФЯЦ-ВНИИЭФ работает в интегрированной библиотечной системе «ИРБИС». В ней предлагается ведение тезауруса в отдельно выделенной базе TEZ. В процессе работы стало очевидно, что применительно к нашей конкретной задачи тезаурус в «ИРБИС» имеет недостаток, так как отображаются не все необходимые семантические связи.



Помогла выйти из сложившейся ситуации статья Прозорова Ивана Евгеньевича «Ведение политематического тезауруса в системе автоматизации библиотек ИРБИС: опыт Центральной городской публичной библиотеки им. В.В.Маяковского». В ней рассматривались проблемы ведения информационно-поискового тезауруса и показан вклад специалистов ГПНТБ в разрешения этой проблемы. В часности, заведующим отделом разработки и совершенствования АБИС ГПНТБ России Александром Иосифовичом Бродовским, было предложено для ведения тезауруса использовать универсальный рубрикатор URUB. Для этого в БД URUB были внесены изменения, а при помощи средств ИРБИС-навигатора «нарисовна необходимая структура тезауруса». Отмечалось, что при помощи специалистов ГПНТБ ведение и отображение тезауруса в БД URUB соответствует требованиями ГОСТа. В конце статьи И.Е.Прозоров ссылался на готовность предоставить служебные файлы, с разрешения А.И.Бродовского, всем заинтересованным специалистам. Мы воспользовались предоставленной возможностью и выражаем ему огромную благодарность.

Нашими специалистами были внесены изменения в файлы программы. Приведены в соответствия с нашими требованиями словари – источник информации, систематический и категорийный индексы. При переводе словарного массива тезауруса в БД URUB, было учтено, что загружаемые нижестоящие термины не группируются по алфавиту, поэтому предварительно словарный массив был обработан.



Вторичные источники лексики

Источником лексики служили ГРНТИ, УДК и специализированный отраслевой классификатор НТИ отражающий наименования объектов техники и технологий, а также наименования элементов объектов техники и технологий Госкорпорации «Росатом».

Дополнялась лексика тезауруса путём обработки дополнительных источников информации:

- Законодательства Российской Федерации в области использования атомной энергии в мирных и оборонных целях:
  - Федеральный закон: N 317-ФЗ от 01.12.2007 «О Государственной корпорации по атомной энергии «Росатом»; N 116-ФЗ от 21.07.1997 «О промышленной безопасности опасных производственных объектов»; N 190-ФЗ от 11.07.2011 «Об обращении с радиоактивными отходами и о внесении изменений в отдельные законодательные акты Российской Федерации».
  - Федеральные нормы и правила в области использования атомной энергии: ОПБ-88/97 «Общие положения обеспечения безопасности атомных станций»; НП-082-07 «Правила ядерной безопасности реакторных установок атомных станций» и др.;
- Статьи Законов (Основные понятия, используемые в настоящем Федеральном законе, Объекты применения настоящего Федерального закона), содержат список терминов, снабжённых примечаниями. Федеральные нормы и правил – кроме списка терминов дополнены списком допустимых сокращений. Допустимые сокращения использовались в процессе создания тезауруса, как синонимы терминов.

- ГОСТ 26392-84. Безопасность ядерная. Термины и определения; ГОСТ Р 52103-2003. Ускорители заряженных частиц. Термины и определения; ГОСТ Р 50996-96. Сбор, хранение, переработка и захоронение радиоактивных отходов. Термины и определения; ГОСТ 22574-77. Материалы ядерные делимые. Термины и определения; ГОСТ 23082-78. Реакторы ядерные. Термины и определения и др.

При анализе ГОСТов выявлено, что помимо списка терминов и допустимых сокращений, в ГОСТе входят термины «Недопустимые к применению» (аскрипторы) и термины которые «Разрешается применять в случаях, исключающих возможность различного толкования».

Дополнялась термины и из специализированных энциклопедий, монографий и учебников. Термины для области «Ядерная техника» частично взяты в издании, опубликованном в Государственном комитете по использованию атомной энергии СССР ЦНИИТЭИатоминформ – «Тезаурус по атомной науке и технике», содержащем около 21тыс. терминов. В нём в качестве источников терминологического наполнения использовался Тезаурус Международной системы ядерной информации (ИНИС), и были учтены терминологические предложения индексаторов и специалистов; а так же издание «Атомная энергетика в терминах» ред. Л.В.Константинов; и в краткой энциклопедии «Атомная энергия» редактор В.С.Емельянов.

В области ядерной медицины мы не нашли справочников и тезаурусов, отражающих узконаправленную технологию, отвечающему современному уровню развития ядерной медицины. Терминология данного направления была выбрана из пособия, выпущенного в 2012 «Физика ядерной медицины» авторы В.Н. Беляев, В.А. Климанов, включающего вопросы: физический фундамент ядерной медицины; устройство и основные характеристики аппаратуры, способы получения основных радиофармпрепаратов; радионуклидную терапию; проблемы радиационной безопасности в ядерной медицине.

Термины для области «Безопасность и аварии» были дополнены из словаря под редакции В.Н. Михайлова «Вопросы безопасности и стратегической стабильности» (2009г.), содержащего около 800 терминов и их определений. В словаре приводится терминология, относящаяся к общим вопросам международной и национальной безопасности, проблемам ядерного сдерживания и стратегической стабильности; вопросам ядерных вооружений, ядерных зарядов и боеприпасов, их созданию, испытаниям и эксплуатации; вопросам нераспространения ядерного оружия и ядерных материалов; приведен перечень договоров и соглашений, касающихся ядерно-оружейной проблематики. Энциклопедия «Ядерное нераспространение» гл. ред. А.В. Хлопков (2009) так же стала источником терминов по ключевым вопросам современного состояния проблем ядерного оружия и режима нераспространения.

Дополнительные источники информации позволяют уточнить значение включенных в тезаурус лексических единиц, и проверить их соответствие отечественным и международным правовым, нормам. Обеспечить однозначность и точность семантики дескрипторов, а также установить семантические отношения между терминами, существующими в понятийной системе отраслевого тезауруса, и отразить их в виде парадигматических отношений между соответствующими дескрипторами.

В Тезаурус были включены следующие типы лексических единиц: одиночные слова – радиоактивность, ядра, нуклид; именные словосочетания – эффект Мессбауэра, интеграл Фурье; лексически значимые компоненты сложных слов; термины описательные обороты – каналы для выпуска излучения из реакторов, отравление продуктами деления, сокращения слов и словосочетаний (СУЗ, ВВЭР, РАО), в словник были также включены соответствующие им полные формы. Сокращения соответствуют требованиям стандартов на сокращения слов в библиографических описаниях. Лексические единицы использовались в Тезаурусе в единственном и во множественном числе.

Определение отношений в тезаурусе

В процесс создания тезауруса самая большая, наиболее кропотливая и трудоёмкая работа – это установление отношений между терминами. Для этого мы использовали предметные указатели различной справочной, энциклопедической и учебной литературы.

Наименьшие вопросы возникали в отношениях между терминами «выше-ниже», объединяющие самые разные типы иерархических отношений: род – вид (электромагнитного излучения –

ультрафиолетовое излучение, целое – часть (ядерный реактор – отражатель нейтронов), сырьё – продукт (уран – ядерное топливо). Отношениями равнозначности (синонимия) – это связь между терминами, различными по звучанию и написанию, но имеющими одинаковое или очень близкое лексическое значение.

Объединением иерархических отношений и синонимии в качестве парадигматических отношений фиксируются ассоциативными отношениями. Допускается включать в ассоциативное отношение все виды отношений, кроме синонимии и отношения род-вид. В Тезаурусе регистрируются те ассоциативные отношения, которые полезны для информационного поиска и представления понятийной системы в целом.

Все отношения создают сложную сеть понятий, и знание о том, где находится понятие в этой сети. Одно и то же понятие может попасть в различные родовые понятия.

Распределение лексических единиц по семантическим категориям

Все лексические единицы, которые были отобраны для включения в Тезаурус, распределялись по семантическим категориям, выявленным на основе анализа области деятельности.

Процессы нераспространение, транспортировка, моделирование, обезвреживание, визуализация.

Свойства надёжность, качество, химические свойства, цвет, прочность, устойчивость, вынужденный, нестационарный .

Предметы томографы, гамма-камеры, ядерные реакторы, компрессоры, насосы, ядерные материалы, радиоактивные вещества, изотопы.

Остальные лексические единицы, включенные в тезаурус, являются аскрипторами или недескрипторами. Они используются для обозначения одного или нескольких близких понятий разными словами, сопряженным с одним и тем же дескриптором.

Вся система отношений в дескрипторной статье фиксирует положение дескриптора в понятийной системе отрасли, представленной в тезаурусе. Каждый дескриптор является как бы центром одного семантического поля, а тезаурус состоит из таких пересекающихся между собой полей. Место дескриптора в этом поле и место этого поля в тезаурусе раскрывают значение дескриптора.

Построение словарных статей

Понятийная структура в отраслевом Тезаурусе отражена в виде совокупности дескрипторов, представляющих эти понятия, и парадигматических связей между дескрипторами. Значение каждого дескриптора, зафиксировано в словарной статье.

Список полей словарной статьи: Термин, Систематический индекс, Категорийный индекс, Примечание, Вышестоящий дискриптор, Ассоциативные дискрипторы, Для недескрипторов Синоним, Для недескрипторов комбинация терминов, Для недескрипторов перечень альтернативных дескрипторов, Нижестоящие дискрипторы, Источник информации. Набор полей для дескриптора и аскриптора отличается.

После построения словарной статьи для каждого дескриптора и аскриптора получается контролируемый словарь, термины которого связаны парадигматическими отношениями и ассоциативной связью.

Таким образом, тезаурус – это терминологический ресурс, реализованный в виде словаря понятий и терминов со связями между ними. Это множество допустимых обозначений для описания содержания документа из некоторой коллекции текстов.

Представление тезауруса на портале НТИ

В настоящее время тезаурус содержит более 7000 лексических единиц, это базисный словарь для информационного массива портала НТИ, сопряженный с тезаурусом МАГАТЭ. Тезаурус успешно интегрирован в портал НТИ и используется в качестве информационно-поискового инструмента, с целью наибольшей эффективности информационного поиска.

Таким образом, лингвистическое обеспечение портала НТИ Госкорпорации «Росатом» предоставляет пользователям терминосреду, позволяющую максимально облегчить поиск научно-технической информации, раскрыть содержание хранящихся источников информации в удобной для пользователя форме.

Список использованных источников

1. ГОСТ 7.25-2001 Тезаурус информационно-поисковый одноязычный. правила разработки, структура, состав и форма представления/ Сборник основных российских стандартов по библиотечно-информационной деятельности. – СПб.: Профессия, 2006. – 547 с.
2. ГОСТ 7.59-2003 Индексирование документов. Общие требования к систематизации и предметизации/ Сборник основных российских стандартов по библиотечно-информационной деятельности . – СПб. : Профессия, 2006. – 547 с.
3. ГОСТ 7.74-98 Межгосударственный рубрикатор научно-технической информации. структура, правила использования и ведения./ Сборник основных российских стандартов по библиотечно-информационной деятельности . – СПб. : Профессия, 2006. – 547 с.
4. Барахнин В.Б. Технология создания тезауруса предметной области на основе предметного указателя энциклопедии /Вычислительные технологии.-2007-Т.12.- Специальный выпуск 2.- с.3-7.
5. Гиляревский Р.С., Шапкин А.В., Белозеров В.Н. Рубрикатор как инструмент информационной навигации.- СПб. : Профессия, 2008.- 352 с.
6. Кушнерук С.П. Тезаурус документальных средств/ Известия ВолгГТУ.
7. Мдивани Р.Р. Тезаурусы ИНИОН РАН по социальным и гуманитарным наукам. // Научно-техническая информация. Сер. Организация и методика информационной работы.- 2013.-№7.- с.23-27.
8. Онтологии и тезаурусы: модели, инструменты, приложения [Текст] : учеб. пособие. – М.: БИНОМ. Лаборатория знаний, 2011. – 173 с.: ил.
9. Оранская, Л. И. Некоторые особенности использования дескрипторного поискового языка в библиографической ИПС универсальной библиотеки / Л. И. Оранская // Научные и технические библиотеки. – 1997.- Вып. 9. – С.13-22
10. Рубашкин В.Ш., Лахути Д.Г. О языке и средствах диалога с экспертом в предметной области в системе ведения семантического словаря. // Научно-техническая информация. Сер.Организация и методика информационной работы.- 2002.-№7.- с.7-15
11. Сукиасян Э.Р. Системный анализ проблем управления качеством информационно-поисковой системы. // Научно-технические библиотеки.-1995.-№3.-с.6-15.
12. Филиппович Ю.Н., Прохоров А.В. Семантика информационных технологий: опыты словарно-тезаурусного описания. М.: Изд-во МГУП, 2002, 368 с.
13. Щербинина Г.С. Философия координатного индексирования. // Научно-технические библиотеки.-2000.-№9.-с.67-78