

УДК 004.912

DOI:10.25729/ESI.2023.31.3.005

## Тезаурус по информационным технологиям в электронной библиотеке ФИЦ ИВТ

Барахнин Владимир Борисович, Федотова Ольга Анатольевна

Федеральный исследовательский центр информационных и вычислительных технологий, Россия, Новосибирск, *of@ict.nsc.ru*

**Аннотация.** Статья посвящена описанию тезауруса по информационным технологиям, разработанному в рамках целевой программы развития информационных ресурсов Отделения «Электронная библиотека Сибирского отделения РАН». Тезаурус активно используется в учебном процессе студентов НГУ, а также для автоматической связи документов в электронной библиотеке, созданной и эксплуатируемой в Федеральном исследовательском центре информационных и вычислительных технологий (ФИЦ ИВТ). Для реализации тезауруса была выбрана схема данных Zthes, главное преимущество которой состоит в ее соответствии модели сетевого протокола Z39.50, что позволяет не только работать с собственным локальным тезаурусом, но и подключать при необходимости тезаурусы, расположенные в сети. В схему данных Zthes разработчиками были добавлены новые элементы. В настоящий момент тезаурус реализован на трех языках (русском, казахском и английском) и содержит свыше 20 000 терминов.

**Ключевые слова:** информационная система, электронная библиотека, словарь-справочник, классификация, информационно-поисковый тезаурус, Z39.50, схема данных, Zthes, метаданные

**Цитирование:** Барахнин В.Б. Тезаурус по информационным технологиям в электронной библиотеке ФИЦ ИВТ / В.Б. Барахнин, О.А. Федотова // Информационные и математические технологии в науке и управлении. – 2023. – № 3(31). – С. 50-59. – DOI:10.25729/ESI.2023.31.3.005.

**Введение.** В 1998 году в Сибирском отделении РАН была сформирована целевая программа развития информационных ресурсов Отделения под общим названием «Электронная библиотека Сибирского отделения РАН».

Для решения проблемы информационной обеспеченности научных исследований начались работы по созданию Интегрированной Распределенной Информационной Системы (ИРИС СО РАН) [1], являющейся полнофункциональной системой информации об интеллектуальном потенциале Сибирского отделения РАН (т.е. информационной системой об институтах, сотрудниках, достижениях и др.) и системой электронной поддержки сбора и накопления научной информации (электронных атласов, электронных коллекций, баз данных и т.п.).

Основные направления программы были связаны с информационной поддержкой научных исследований, проводимых в Отделении, а также с формированием и развитием собственных электронных ресурсов по основным отраслям наук, с управлением этими ресурсами, с созданием и поддержкой электронных коллекций, публикаций и библиографических ресурсов.

Несмотря на определенные противодействия, благодаря поддержке академиков Ю.И. Шокина, В.К. Шумного и И.Ю. Корапачинского были созданы первые коллекции.

На сегодняшний день наиболее эффективным способом решения проблем организации доступа к распределенным информационным ресурсам является организация информации о них в системах, облаченных в форму электронных библиотек. В рамках нашего подхода электронные библиотеки рассматриваются как средства поддержки отдельной конкретной технологии работы с цифровой информацией, образующие новый класс информационных систем (ИС), предназначенных для управления информационными ресурсами [2].

Под термином электронная библиотека (ЭБ) в настоящей работе понимается система управления структурированными каталогизированными коллекциями разнородных электронных (цифровых) объектов (ресурсов). ЭБ не только обеспечивает многосторонний поиск и навигацию по каталогам (в отличие от печатных изданий, микрофильмов и других носителей),

но и непосредственно предоставляет пользователю найденный ресурс (публикацию, документ, фотографию, описание факта и др.), а также дополнительные сведения о нем, например, географическую привязку, информацию об авторах, информацию о фактах, библиографию, перечень организаций и т.д.

В работе [3] представлены основные принципы реализации ИРИС, основанные на использовании идеи ЭБ, а в [4] определены основные цели, стоящие перед ЭБ:

- управление информационными ресурсами;
- обеспечение и управление доступом к информации;
- долговременное хранение информации;
- сохранение научного и культурного наследия;
- поддержка аналитической работы с информацией;
- повышение эффективности научных исследований и обучения.

Исходя из целей ЭБ и анализа существующих систем, направленных на поддержку научных исследований, в работе [5] сформулированы следующие функциональные требования к модели ЭБ:

- надежное долговременное и защищенное от исчезновения хранение информации;
- актуальность, полнота, достоверность происхождения документов;
- географическая привязка информации;
- наличие большого числа словарей-классификаторов (справочников) для обеспечения идентификации и классификации ресурсов;
- поддержка неоднородных и слабо структурированных информационных ресурсов;
- поддержка взаимосвязей информационных ресурсов;
- наличие интеллектуальных служб обслуживания запросов пользователя;
- наличие программных интерфейсов для поддержки аналитической работы пользователя с помощью программных приложений;
- поддержка требований интероперабельности как на программном, так и на семантическом уровне;
- поддержка работы с внешними источниками.

С учетом целей и требований была разработана модель ЭБ [6, 7, 8]. Основным выводом стало то, что она должна быть многоуровневой и состоять как минимум из следующих компонентов: распределенное хранилище данных – репозиторий, сервер метаданных, сервер приложений, словари-справочники (рис. 1).

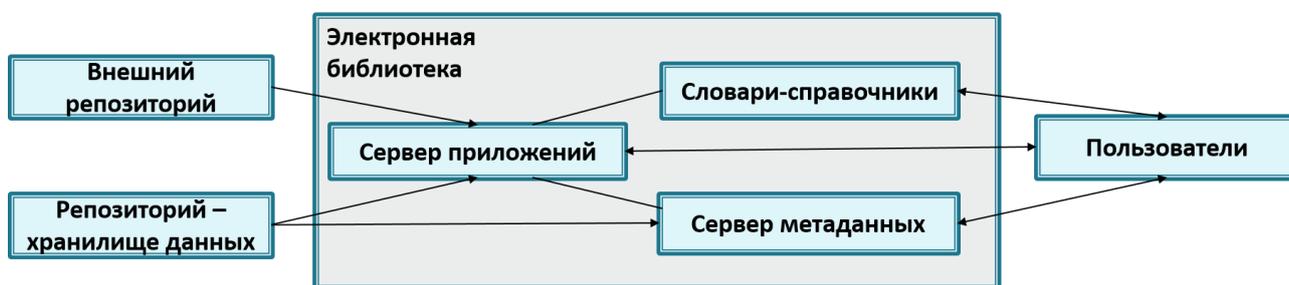


Рис. 1. Архитектура ЭБ

Разработанная модель ЭБ была реализована в виде Системы Управления Электронными Библиотеками (СУЭБ ИРИС) и эксплуатируется в ФИЦ ИВТ с 2004 года.

В работе использовались СУБД MySQL и PostgreSQL. Поддержка интероперабельности обеспечивается использованием открытых стандартов. Основные интерфейсы ЭБ (шаблоны представления документов библиотеки, формирование списков документов, агенты, определяющие связи документов и т. п.) написаны на языке PHP. Для осуществления логического и

визуального контроля ввода значений некоторых элементов метаданных написаны процедуры на языке Java Script. Для обработки регулярных выражений для задач поиска было принято решение об использовании поисковой машины протокола Z39.50, входящей в состав СУЭБ.

На сегодняшний день ЭБ включает свыше 60 коллекций. Среди них: научно-образовательный ресурс «Вычислительные системы» (свыше 1500 статей), ЭБ учебных пособий по курсам «Современные проблемы информатики и вычислительной техники», «Информатика», «Экология» и др., ЭБ по моделям динамики биосферы, по теплофизическим свойствам и фазовым превращений теплоносителей и др., тезаурусы по «Информационным технологиям» (свыше 20000 терминов), «Горнорудной терминологии» и др., персоны (около 800 записей), коллекции учебных курсов по информатике и полные тексты публикаций, а также контролируемые словари и классификаторы. На рисунке 2 приведен пример из электронного ресурса «Вычислительные системы», иллюстрирующий материалы о наследии Джона фон Неймана: архитектура ЭВМ, биография, ссылки на репозиторий. Справа представлена словарная статья с ключевыми словами, ссылки на литературу и персону, слева – фрагменты биографии и ссылки на репозиторий.

**Архитектура фон Неймана**

**Синонимами:** Неймановская архитектура; Принстонская архитектура;

В статье фон Неймана «First Draft of a Report on the EDVAC» определены основные устройства ЭВМ, с помощью которых реализованы его принципы. Большинство современных ЭВМ по своей структуре отвечают принципу программной неймановской ЭВМ: содержит: память, устройство управления, арифметико-логическое устройство и устройство ввода-вывода.

**Ключевые термины, связанные с термином "Архитектура фон Неймана":**

1. Архитектура вычислительной машины
2. Вычислительная машина с непосредственными связями
3. Гарвардская архитектура
4. Машина Тьюринга
5. Принципы фон Неймана
6. Процедурный язык
7. Управление последовательностью команд

**Литература**

**Дополнительная:**

1. John von Neumann First Draft of a Report on the EDVAC // Journal IEEE Annals of the History of Computing October 1993 Page 27-75
2. Черняк Л. Архитектура фон Неймана, реконфигурируемые компьютерные системы и антимашина // Открытые системы. Информатика. 1993. № 1. С. 10-15.
3. Цильбер Б.Я. Организация ЭВМ и систем : Учебник для вузов / Б.Я. Цильбер, С.А. Орлов - 2-е изд., стереотип. - М.: Мир, 1998. - 978-5-49807-862-5.

**Ссылки на персон:**

1. Нейман Джон фон

**Ключевые термины (головные):** Архитектура вычислительной машины; Принципы фон Неймана;

● **Контекстный поиск:** Задайте образец для поиска:

---

**Джон фон Нейман**  
28 декабря 1903, Будапешт — 8 февраля 1957, Вашингтон

**Основные достижения:**

Джон фон Нейман (John von Neumann) сделал важный вклад в квантовую физику, квантовую логику, функциональный анализ, теорию множеств, теорию автоматов, кибернетику, информатику, экономику и другие отрасли науки.

Наиболее известен как создатель современной архитектуры компьютеров (архитектура фон Неймана), применением теории операторов к квантовой механике (алгебра фон Неймана), а также как участник Манхэттенского проекта и как создатель теории игр и концепции клеточных автоматов.

В 1946 году Джон фон Нейман на основе критического анализа конструкции ENIAC предложил ряд новых идей организации ЭВМ, в том числе концепцию хранимой в памяти программы. В результате реализации идей фон Неймана была создана архитектура ЭВМ, во многих чертах сохранившаяся до настоящего времени.

В 1951 году Джон фон Нейман, используя идеи Эккерта формулирует принципы организации вычислительных систем (Принципы фон Неймана), которые легли в основу всех современных компьютеров:

1. Универсальная вычислительная машина должна содержать несколько основных компонент, таких как арифметическое устройство, память, устройство управления и связи с оператором.
2. Вычислительная машина должна быть полностью автоматической, т.е. после начала вычисления, работа машины не должна зависеть от человека-оператора.

● John von Neumann First Draft of a Report on the EDVAC // Journal IEEE Annals of the History of Computing archive Volume 15 Issue 4, October 1993 Page 27-75

The first draft of a report on the EDVAC written by John von Neumann is presented. This first draft contains a wealth of information, and it had a pervasive influence when it was first written. Most prominently, Alan Turing cites it in his proposal for Pilot automatic computing engine (ACE) as the definitive source for understanding the nature and design of a general-purpose digital computer.

Ссылка на адрес в репозитории: First Draft of a Report on the EDVAC  
URL-адрес источника ресурса: First Draft of a Report on the EDVAC  
Ссылка на репозиторий DSpace: First Draft of a Report on the EDVAC

Ссылки на авторов публикации:  
Нейман, Джон фон

Рис. 2. Пример электронного ресурса «Вычислительные системы»

В 2015 году для поддержки этих научно-образовательных ресурсов был создан трехязычный тезаурус предметной области «Информатика», содержащий термины на русском, английском и казахском языках (21672 термина). В качестве платформы реализации тезауруса была использована вышеуказанная СУЭБ. Тезаурус активно используется в учебном процессе студентов НГУ, а также для автоматической связи документов в электронной библиотеке.

**1. Проблемы классификации.** Стандартным подходом к систематизации информации является классификация документов с помощью таксономий, т.е. предметной классификации, которая группирует термины в виде управляемого словаря (тезауруса) и упорядочивает их (словари) в виде иерархических структур. Для описания какой-либо предметной области обычно используется определенный набор ключевых терминов, каждый из которых обозначает или описывает какое-либо понятие из данной предметной области [9]. Основу классификации составляет выделение понятий (ключевых терминов), установление парадигматических отношений (например, типа родитель – потомок) между ними и сопоставление анализируемого документа выделенным понятиям.

Основная проблема состоит в высокой трудоемкости и стоимости ручного составления тезауруса, малой гибкости процесса его построения. В тезаурусах для ручного индексирования совокупности близких понятий сводятся к одному, наиболее представительному понятию для уменьшения субъективности индексирования. В автоматизированных тезаурусах семантически близкие понятия представлены в виде отдельных единиц, что позволяет использовать при поиске синонимические ряды. Трудность построения тезауруса, соответствующего всему тематическому многообразию индексируемой информации, является основной причиной его непопулярности в современных информационных системах. Но если рассматривать вопрос об эффективности информационных систем в определенных областях знаний, то создание и использование специализированных тезаурусов в таких системах представляет несомненный интерес и переводит систему в совершенно другой качественный класс.

Эффективность информационно-поисковых систем для поддержки научно-образовательной деятельности напрямую зависит от использования специализированных тезаурусов. Дополнительной особенностью при создании информационной системы поддержки в научно-образовательной деятельности для таких стран, как Казахстан или Россия, является необходимость поддерживать поисковые и классификационные процессы одновременно на нескольких языках: для России, в основном, на двух языках (русском и английском), а для Казахстана – как минимум на трех (русском, английском и казахском) языках. Следовательно, документы должны индексироваться в трех различных пространствах, соответствующих трем языкам, между элементами которых должны быть установлены отношения эквивалентности, либо в интегрированном пространстве, который задается многоязычным тезаурусом.

Отметим, что элементы признаков пространств могут быть представлены в документе в различных словоформах, поэтому важнейшей проблемой является учет морфологии конкретного языка при индексации документов.

**2. Тезаурусы в информационном поиске.** Согласно определению международной организации по стандартизации (ISO), тезаурус является словарем, управляемым языком индексации, формально организованным для того, чтобы установить явные априорные отношения между понятиями [10, 11]. Это определение устанавливает лексические единицы и семантические отношения между этими единицами в качестве элементов, составляющих тезаурус. Тезаурусные отношения (род – вид, часть – целое и др.) налагаются на структуру таксономии, т.е. идентифицируются основные таксономии предметной области.

Исторически тезаурусы создавались для ручного индексирования документов и при их создании не принимались во внимание вопросы, связанные с автоматической индексацией.

Один из первых в истории и наиболее известных на сегодня тезаурусов (идеографических словарей) составлен британским лексикографом Питером Марком Роже и опубликован в 1852 г. Оригинальное название тезауруса Роже – *Thesaurus of English Words and Phrases* (Тезаурус английских слов и фраз).

В области машинного перевода впервые тезаурус был использован М. Мастерман (M. Masterman). В 1961 г. появилась работа [12], в которой определялся базовый словарь для 15000 понятий. При помощи тезаурусов устанавливалось соответствие между языком запросов пользователя и документами в информационной системе.

Ю.А. Шрейдер [13] предлагал рассматривать тезаурус как систему знаний, отраженных языком, тогда тезаурус становится интересным сам по себе, а не только как вспомогательный инструмент.

Среди универсальных автоматизированных тезаурусов следует отметить интеллектуальный электронный тезаурус для английского языка WORDNET, разработанный в Принстон-

ском университете, аналогичный тезаурус RussNet для русского языка, разрабатываемый сотрудниками кафедры математической лингвистики Санкт-Петербургского государственного университета, многоязычный тезаурус ЮНЕСКО.

Среди специализированных тезаурусов, реализованных уже на машинных носителях с программной поддержкой для пользователей, можно отметить следующие: многоязычный тезаурус EuroVOC, специально разработанный для индексирования и поиска информации в документальных и библиографических базах данных Европейского Союза; тезаурусы и контролируемые словари Исследовательской службы Библиотеки Конгресса США; тезаурусы Американского общества индексирования ASI; известный тезаурус AGROVOC, охватывающий терминологию сельского и лесного хозяйства; компьютеризированный тезаурус медицинской терминологии SNOMED и др.

Среди российских тезаурусов следует отметить тезаурус парламентской библиотеки РФ, разработанный как средство индексирования и поиска документов в автоматизированной информационно-библиотечной системе.

**3. Тезаурусы в описании информации.** Информационно-поисковый тезаурус (в соответствии с определениями стандартов) – это нормативный (контролируемый) словарь ключевых терминов на естественном языке с явно указанными семантическими отношениями между терминами, предназначенный для описания содержания документов и поисковых запросов [14]. Тезаурус предназначен для описания конкретной предметной области, каждый термин которой обозначает или описывает какое-либо понятие из данной предметной области.

Термины, составляющие тезаурус, подразделяются на дескрипторы (авторизованные термины) и недескрипторы (аскрипторы). Дескрипторы однозначно соответствуют понятиям предметной области. Отношения между дескрипторами обычно разделяются на два типа: иерархические и ассоциативные. Иерархические отношения обычно рассматриваются как несимметричные и транзитивные.

В соответствии с ГОСТ 7.25-2001 [15] иерархические отношения обладают свойствами транзитивности и антисимметричности, которые могут быть использованы при избыточном индексировании в интересах повышения эффективности информационного поиска.

Применяемые в информационно-поисковых тезаурусах иерархические отношения могут дифференцироваться на отдельные виды. Основным иерархическим отношением, используемым в информационно-поисковых тезаурусах, является родовидовое отношение (родитель – потомок). Родовидовая связь устанавливается между двумя дескрипторами, если объем понятия нижестоящего дескриптора входит в объем понятия вышестоящего дескриптора. Также в качестве иерархического отношения в информационно-поисковых тезаурусах может устанавливаться отношение «часть – целое». Отношение ассоциации является неиерархическим и ассоциативным. Основное назначение установления ассоциативных отношений между дескрипторами информационно-поискового тезауруса – указание на дополнительные дескрипторы, полезные при индексировании или поиске [16].

Отметим, что описанная в национальных и международных стандартах модель информационно-поискового тезауруса предназначена для его использования в процессе ручного, экспертного анализа документов.

Главной целью разработки традиционных информационно-поисковых тезаурусов является использование их единиц (дескрипторов) для описания основных тем документов в процессе ручного индексирования. Поэтому важно, чтобы набор дескрипторов информационно-поискового тезауруса позволял описывать тематику документов предметной области. При этом сам процесс индексирования по такому тезаурусу базируется на лингвистических, грамматических знаниях, а также знаниях о предметной области, которые имеются у профессио-

нальных индексов текстов. Индексатор сначала должен прочитать текст, понять его и затем изложить содержание текста, пользуясь дескрипторами, указанными в информационно-поисковом тезаурусе. Индексатор должен хорошо понимать всю терминологию, использованную в тексте, тогда для описания основной темы текста ему понадобится значительно меньшее количество терминов [9, 16].

**4. Схема данных Zthes.** Для реализации тезауруса была выбрана схема данных Zthes, главное преимущество которой состоит в ее соответствии модели сетевого протокола Z39.50, что позволяет не только работать с собственным локальным тезаурусом, но и подключать при необходимости тезаурусы, расположенные в сети [17].

Исходя из общей идеологии Z39.50, доступ к любой базе данных должен осуществляться через единую стандартную схему данных, на которую должны быть корректно отображены все частные структуры. Схема Zthes впервые была описана в стандарте ISO 25964-1:2011 [14]. Стандарт предназначен для использования в приложениях для информационного поиска. Он применим в отношении словарей, используемых для поиска и извлечения информации из информационных ресурсов всех видов, независимо от вида носителей (текст, звук, фотографии и видеозаписи, физические или мультимедийные объекты), включая базы знаний и порталы, библиографические базы данных, текстовые, музейные и мультимедийные коллекции и содержащиеся в них объекты. Стандарт используется в отношении одно- и многоязычных тезаурусов. Последняя версия стандарта утверждена в 2022 году.

Схема Zthes предполагает использование весьма ограниченного набора типов отношений между терминами. Это сделано преднамеренно для большей совместимости.

Между терминами, в соответствии с рекомендациями стандарта, устанавливаются следующие типы связей:

- *BT* – связь с родительским термином, т. е. с термином более широкого смысла;
- *NT* – связь с дочерним термином, т. е. с термином более узкого смысла; связь *BT↔NT* является взаимно-обратной;
- *USE* – связь с термином, который используется вместо этого;
- *UF* – взаимно-обратная связь *USE*, *USE↔UF*;
- *RT* – связь, определяющая связанный термин;
- *LE* – связь между лингвистически эквивалентными терминами.

Связи *BT* и *NT*, а также *USE* и *UF*, взаимно-обратные. Связи *RT* и *LE* симметричны.

Кроме того, тип термина определяется в соответствии с рекомендациями стандарта.

В схеме Zthes выделяются следующие типы терминов:

- *TT* – термин верхнего уровня, т.е. термин, не имеющий связанных терминов более широкого класса (терминов с типом связей *BT*);
- *NT* – термин не верхнего уровня, т. е. дескриптор, имеющий связи типа *BT*;
- *ND* – неосновной термин;
- *NL* – фиктивный термин, т. е. термин, не используемый для индексации документов, но включенный в иерархию, чтобы указать логический базис раздела классов.

**5. Реализация тезауруса.** Практическая реализация тезауруса (алгоритмы, программы и заполнение базы данных), основанная на описанной выше схеме данных Zthes, была выполнена авторами настоящей работы, а также аспирантами факультета информационных технологий (ФИТ) Новосибирского государственного университета (НГУ), под руководством проф. А.М. Федотова и при его непосредственном участии.

Как было отмечено в предыдущем разделе, схема Zthes использует ограниченный набор данных для совместимости. В тезаурус авторами были добавлены расширения, связанные с удобством программной реализации:

- termNormName – нормальная форма термина (единственное число, именительный падеж и т.п.), элемент для учета морфологии, в частности, в алгоритмах поиска;
- termLinkID – алгоритмически вычисляемый уникальный идентификатор, характеризующий термин, обеспечивает ряд удобств при программировании;
- termScoreNote – элемент, содержащий словарную статью (или ссылку на нее) с подробным объяснением понятия, уточняющего смысл и область применения, ввиду ориентации разработанного тезауруса на использование в образовательном процессе.

Контролируемый словарь типа связей (relationType) был расширен признаком ‘SYN’ – полный синоним, куда заносится понятийный эквивалент термина, получаемый, как правило, из различий в его написании (например, вычислительные системы ≡ выч. системы).

Созданные компоненты реализации тезауруса (рис. 3) позволяют просматривать, редактировать и добавлять термины тезауруса в систему через web-формы, а также импортировать и экспортировать термины в виде XML-, RDF-, DTD-файлов. Преобразование данных в схемы данных SKOS, MARC, MODS и представление терминов в браузере производится с помощью XSLT-преобразования.

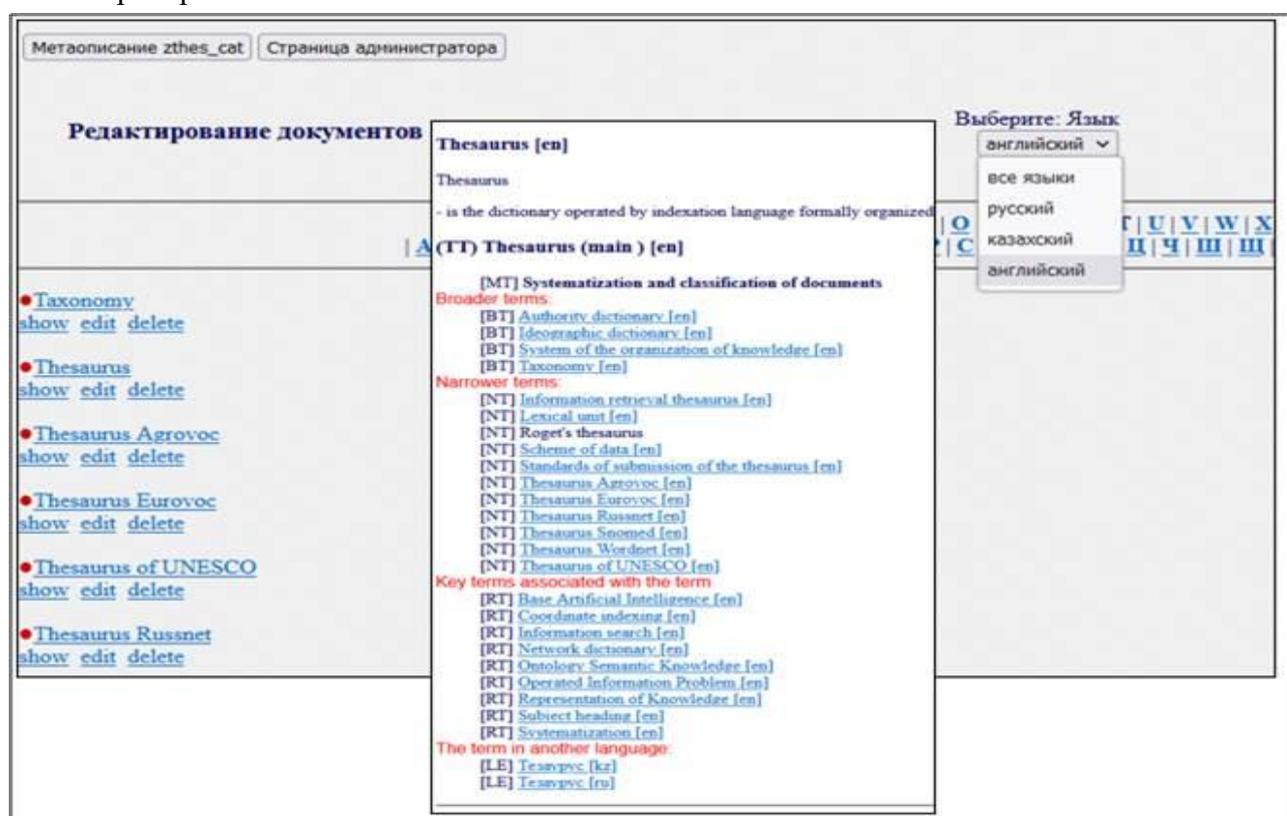


Рис. 3. Иллюстрация редактирования тезауруса

Базовое метаописание термина включает следующие элементы и связи: название термина, UID (основной link\_id), описание термина, язык термина, связь с родительским термином, связь с дочерним термином, термин, который используется вместо данного, аскрипторы, полные синонимы, коды УДК, коды классификатора ГРНТИ и др.

**Заключение.** На основе анализа стандартов и различных подходов к реализации тезаурусов принято решение об использовании схемы данных Zthes для создания тезауруса по информатике. В настоящий момент тезаурус реализован на трех языках (русском, казахском и английском) и содержит около 22 000 терминов. Использование тезауруса в электронных библиотеках наиболее эффективно при постоянной модернизации тезауруса, его интеграции в базу данных и соответствующем уровне тематической специализации. Если изначально тезаурус

урус использовался только для навигации по ресурсам библиотеки, классификации и рубрикации, то теперь он активно применяется в учебном процессе студентов ФИТ НГУ и может рассматриваться как самостоятельный словарь-справочник.

#### Список источников

1. Шокин Ю.И. Эволюция информационных систем: от Web-сайтов до систем управления информационными ресурсами / Ю.И. Шокин, А.М. Федотов, О.Л. Жижимов, О.А. Федотова // Вестник НГУ. Серия: Информационные технологии, 2015. – Т. 13. – № 1. – С. 117-134.
2. Жижимов О.Л. Некоторые заметки об эволюции цифровых репозиторий традиционных библиотек к полнофункциональным электронным библиотекам / О.Л. Жижимов, Н.А. Мазов, А.М. Федотов // Вестник Владивостокского государственного университета экономики и сервиса. Территория новых возможностей, 2010. – Т. 7. – № 3. – С. 55-63.
3. Шокин Ю.И. Поддержка и развитие распределенных информационно-вычислительных ресурсов в СО РАН / Ю.И. Шокин, А.М. Федотов // Вычислительные технологии. Вестн. КазНУ им. аль-Фараби. Серия: Математика, механика, информатика, 2004. – Т. 42. – Ч. 4. – № 3. – С. 324-334.
4. Федотов А.М. Технология создания корпоративных информационных систем учета трудов научных работников / А.М. Федотов, В.Б. Барахнин, О.Л. Жижимов, О.А. Федотова // Вестник Новосибирского государственного университета. Серия: Информационные технологии, 2011. – Т. 9. – № 2. – С. 31-41.
5. Федотова О.А. Требования к информационной модели электронной библиотеки по научному наследию / О.А. Федотова // Zbornik radova konferencije MIT 2013. – Beograd, 2014. – С. 141-149.
6. Федотов А.М. Модель информационной системы для поддержки научно-педагогической деятельности / А.М. Федотов, В.Б. Барахнин, О.Л. Жижимов, О.А. Федотова // Вестник НГУ. Серия: Информационные технологии, 2014. – Т. 12. – Вып. 1. – С. 89-101.
7. Федотов А.М. Концептуальная модель научно-образовательной информационной системы / А.М. Федотов, М.Н. Абделиева, А.Т. Байдавлетов и др. // Вестник НГУ. Серия: Информационные технологии, 2015. – Т. 13. – № 3. – С. 89-104.
8. Varanov A., Tussupov J., Fedotov A. et al. The creation of information model of digital library for supporting scientific and educational activity. information technology in industry, 2019, vol. 7, no. 2, pp. 41-46.
9. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. / Н.В. Лукашевич. – М.: Изд-во МГУ, 2011. – 512 с.
10. ISO 5964:1985. Documentation – Guidelines for the establishment and development of multilingual thesauri. Geneva: International Organization for Standardization, 1985.
11. ISO 2788:1986. Documentation – Guidelines for the establishment and development of monolingual thesauri. 2nd ed. Geneva: International Organization for Standardization, 1986.
12. Masterman M. Semantic message detection for machine translation, using an interlingua. Proc. international conf. on machine translation, 1961, pp. 438-475.
13. Шрейдер Ю.А. О количественных характеристиках семантической информации / Ю.А. Шрейдер // НТИ: Сер. 2, 1963. – № 10. – С. 35-39.
14. ISO 25964-1:2011. Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval, 2011.
15. ГОСТ 7.25-2001. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления (Система стандартов по информации, библиотечному и издательскому делу) / Межгосударственный совет по стандартизации, метрологии и сертификации. – М.: Стандартинформ, 2002.
16. ISO/IEC 13250:2003. Information technology – SGML Applications – Topic Maps, 2003.
17. Жижимов О.Л. Технологическая платформа массовой интеграции гетерогенных данных / О.Л. Жижимов, А.М. Федотов, Ю.И. Шокин // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии, 2013. – Т. 11. – № 1. – С. 24-41.

**Барахнин Владимир Борисович.** Доктор технических наук, ведущий научный сотрудник Федерального исследовательского центра информационных и вычислительных технологий. ORCID: 0000-0003-3299-0507, WOS: A-5856-2014, AuthorID: 10510, bar@ict.nsc.ru, 630090, г. Новосибирск, пр. Академика Лаврентьева, 6.

**Федотова Ольга Анатольевна.** Младший научный сотрудник Федерального исследовательского центра информационных и вычислительных технологий, ORCID: 0000-0001-5012-2426, ResearcherID: I-9755-2018, of@ict.nsc.ru, 630090, г. Новосибирск, пр. Академика Лаврентьева, 6.

## Thesaurus for information technologies in the digital library of FRC ICT

Vladimir B. Barakhnin, Olga A. Fedotova

Federal Research Center for Information and Computational Technologies,  
Russia, Novosibirsk, *of@ict.nsc.ru*

**Abstract.** The article is devoted to the description of the thesaurus on information technologies developed within the framework of the target program for the development of information resources of the Department "Electronic Library of the Siberian Branch of the Russian Academy of Sciences". The thesaurus is actively used in the educational process of NSU students, as well as for automatic communication of documents in the electronic library created and operated at the FRC ICT. To implement the thesaurus, the Zthes data scheme was chosen, the main advantage of which is its compliance with the Z39.50 network protocol model, which allows not only to work with its own local thesaurus, but also to connect, if necessary, thesauri located on the network. New elements have been added to the Zthes data schema by the developers. Currently, the thesaurus is implemented in three languages (Russian, Kazakh and English) and contains over 20,000 terms.

**Keywords:** information system, digital library, glossary, classification, information retrieval thesaurus, Z39.50, data schema, Zthes, metadata

### References

1. Shokin Yu.I., Fedotov A.M., Zhizhimov O.L., Fedotova O.A. Evolyutsiya informatsionnykh sistem: ot Web-saytov do sistem upravleniya informatsionnymi resursami [The evolution of information systems: from websites to information resource management systems]. Vestnik NGU. Seriya: Informatsionnyye tekhnologii [Vestnik NSU. Series: Information Technologies], 2015, v. 13, no. 1, pp. 117-134.
2. Zhizhimov O.L., Mazov N.A., Fedotov A.M. Nekotoryye zametki ob evolyutsii tsifrovyykh repozitoriyev traditsionnykh bibliotek k polnofunktsional'nym elektronnykh bibliotekam [Some notes about evolution of digital repositories of traditional libraries to full-function electronic libraries]. Vestnik Vladivostokskogo gosudarstvennogo universiteta ekonomiki i servisa. Territoriya novyykh vozmozhnostey [Territory of new opportunities. Bulletin of Vladivostok State University of Economics and Service], 2010, v. 7, no. 3, pp. 55-63.
3. Shokin Yu.I., Fedotov A.M. Podderzhka i razvitiye raspredelennykh informatsionno-vychislitel'nykh resursov v SO RAN [Technologies for designing of distributed information systems to support research]. Vychislitel'nyye tekhnologii. (Sovmestnyy vypusk). Vestn. KazNU im. al'-Farabi. Seriya: Matematika, mekhanika, informatika [Computational technologies. Herald. KazNU named after. al-Farabi. Series: Mathematics, mechanics, computer science], 2004, v. 42, part 4, no. 3, pp. 324-334.
4. Fedotov A.M., Barakhnin V.B., Zhizhimov O.L., Fedotova O.A. Tekhnologiya sozdaniya korporativnykh informatsionnykh sistem ucheta trudov nauchnykh rabotnikov [The technology of creation of corporate information systems for accounting resources, created by researchers]. Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnyye tekhnologii [Vestnik NSU. Series: Information Technologies], 2011, v. 9, no. 2, pp. 31-41.
5. Fedotova O.A. Trebovaniya k informatsionnoy modeli elektronnoy biblioteki po nauchnomu naslediyu [Requirements for the information model of an electronic library on scientific heritage]. Zbornik radova konferencije MIT 2013. Beograd, 2014, pp. 141-149.
6. Fedotov A.M., Barakhnin V.B., Zhizhimov O.L., Fedotova O.A. Model' informatsionnoy sistemy dlya podderzhki nauchno-pedagogicheskoy deyatel'nosti [Model of an information system to support scientific and pedagogical activities]. Vestnik NGU. Seriya: Informatsionnyye tekhnologii [Vestnik NSU. Series: Information Technologies], 2014, v. 12, no. 1, pp. 89-101.
7. Fedotov A.M., Abdelyeva M.N., Baydavletov A.T., Bapanov A.A., Sambetbayeva M.A., Fedotova O.A. Kontseptual'naya model' nauchno-obrazovatel'noy informatsionnoy sistemy [Conceptual model of scientific and educational information system]. Vestnik NGU. Seriya: Informatsionnyye tekhnologii [Vestnik NSU. Series: Information Technologies], 2015, v. 13, no. 3, pp. 89-104.
8. Bapanov A., Tussupov J., Fedotov A. et al. The creation of information model of digital library for supporting scientific and educational activity. Information technology in industry, 2019, v. 7, no. 2, pp. 41-46.
9. Lukashevich N.V. Tezaurusy v zadachakh informatsionnogo poiska [Thesauruses in information retrieval problems]. M., Izd-vo MGU [Publishing house of MSU], 2011, 512 p.
10. ISO 5964:1985. Documentation – Guidelines for the establishment and development of multilingual thesauri. Geneva: International Organization for Standardization, 1985.

11. ISO 2788:1986. Documentation – Guidelines for the establishment and development of monolingual thesauri. 2nd ed. Geneva: International Organization for Standardization, 1986.
12. Masterman M. Semantic message detection for machine translation, using an interlingua. Proc. international conf. on machine translation, 1961, pp. 438-475.
13. Shreyder YU.A. O kolichestvennykh kharakteristikakh semanticheskoy informatsii [On the quantitative characteristics of semantic information]. NTI: Ser. 2, 1963, no. 10, pp. 35-39.
14. ISO 25964-1:2011. Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval, 2011.
15. GOST 7.25-2001. Tezaurus informatsionno-poiskovyy odnoyazychnyy. Pravila razrabotki, struktura, sostav i forma predstavleniya (Sistema standartov po informatsii, bibliotechnomu i izdatel'skomu delu) / Mezhdgosudarstvennyy sovet po standartizatsii, metrologii i sertifikatsii [Monolingual information retrieval thesaurus. Development rules, structure, composition and form of presentation (System of standards for information, library and publishing) / Interstate Council for Standardization, Metrology and Certification]. M., Standartinform, 2002.
16. ISO/IEC 13250:2003. Informatsionnyye tekhnologii. Primeneniye tipovogo obobshchennogo yazyka razmetki. Tematicheskkiye plany [Information technology.SGML Applications. Topic Maps (SGML)], 2003.
17. Zhizhimov O.L., Fedotov A.M., Shokin Yu.I. Tekhnologicheskaya platforma massovoy integratsii geterogennykh dannykh [Technology platform for the mass integration of heterogeneous data]. Vestn. Novosib. gos. un-ta. Seriya: Informatsionnyye tekhnologii [Vestnik NSU. Series: Information Technologies], 2013, v. 11, no. 1, pp. 24-41.

**Barakhnin Vladimir Borisovich.** Doctor of technical sciences, leading researcher, Federal research center for information and computational technologies. ORCID: 0000-0003-3299-0507, WOS: A-5856-2014, AuthorID: 10510, bar@ict.nsc.ru, 630090, Russia, Novosibirsk, Academician Lavrentiev av., 6.

**Fedotova Olga Anatolyevna.** Junior researcher, Federal research center for information and computational technologies. ORCID: 0000-0001-5012-2426, ResearcherID: I-9755-2018, of@ict.nsc.ru, 630090, Russia, Novosibirsk, Academician Lavrentiev av., 6.

Статья поступила в редакцию 31.07.2023; одобрена после рецензирования 01.09.2023; принята к публикации 18.09.2023.

The article was submitted 07/31/2023; approved after reviewing 09/01/2023; accepted for publication 09/18/2023.