ТЕЗАУРУСЫ И ОНТОЛОГИИ В НАУЧНО-ОБРАЗОВАТЕЛЬНЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

-

Федотов А.М.[1,2], Самбетбаева М.А.[2], Федотова О.А.[2,3]

[1] Институт вычислительных технологий СО РАН, г. Новосибирск, Россия
[2] Новосибирский государственный университет, г. Новосибирск, Россия
[3] Государственная научно-техническая библиотека СО РАН, г. Новосибирск, Россия

Аннотация: Статья посвящена обсуждению проблем поиска информации, систематизации и каталогизации «информации и знаний» в современной информационной среде, анализируются этапы развития научной мысли, обусловившие возникновение концепции теории тезауруса с целью выявления комплекса идей для улучшения качества поиска с помощью тезаурусов и онтологий в информационной системе для поддержки научных исследований.

Ключевые слова: Идеографический словарь, управляемый словарь, тезаурус, онтология, систематизация и классификация ресурсов

THESAURUSES AND ONTOLOGIES IN SCIENTIFIC AND EDUCATIONAL INFORMATION SYSTEMS

Fedotov A.M.[1,2], Sambetbaeva M.A.[2], Fedotova O.A.[2,3]

[1] Institute of Computational Technologies SB RAS, .Novosibirsk, Russian Federation
[2] Novosibirsk State University Novosibirsk, Russian Federation
[3] State Public Scientific and Technical Library SB RAS, Novosibirsk, Russian Federation

Abstract: Article is devoted to discussion of problems of information search, systematization and cataloguing of "information and knowledge" in the modern information environment, the stages of development of a scientific thought which have caused emergence of the concept of the theory of the thesaurus for the purpose of identification of a complex of the ideas for improvement of quality of search by means of thesauruses and ontologies in an information system for support of scientific research are analyzed.

Keywords: ideographic dictionary, controlled dictionary, thesaurus, ontology, systematization and classification of resources

Введение. Одним из основных результатов созидательной, социальной и интеллектуальной человеческой деятельности является создание и накопление информационных ресурсов с целью их дальнейшего использования и недопущения утраты опыта предыдущих поколений. Не будет преувеличением сказать, что состояние технологий накопления информации и эффективности их использования значительно влияет на уровень развития производительных сил. Утеря информации приводила к отбрасыванию цивилизации на века назад. Чтобы эффективно пользоваться имеющейся информацией, необходимы инструменты и технологии, при помощи которых могут быть реализованы специальные приемы работы с информацией. Одним из таких приемов является классификация информации [1].

Проблема поиска информации. Проблема поиска информации является одной из вечных проблем, возникающих в человеческой деятельности. Чтобы решить эту проблему, человечество создало «библиотеки» – универсальную систему хранения ин-

формации и знаний, их систематизации и каталогизации [2, 3]. Любой производственный или научный процесс порождает огромные объемы данных, и работать с ними становится все сложнее по мере того, как гигабайты данных превращаются в терабайты. Количество данных когда-нибудь превысит способность компьютеров их обрабатывать, поэтому необходимы новые инструментальные средства и алгоритмы для анализа этих данных. Вместе с тем предъявляются серьезные требования к обеспечению прозрачного доступа и долговременной сохранности «информации». А в результате вопросы «что хранить», «как хранить» и «как найти» остаются самыми существенными: без ответа на них все остальные теряют актуальность.

Нынешнюю технологическую революцию характеризуют не столько сами знания и технологии, сколько применение знаний и информации к генерированию новых знаний и созданию систем, обрабатывающих информацию и осуществляющих ее передачу.

Современные информационные технологии предоставляют исследователю мощный аппарат для манипулирования данными (как набором битов), но не информацией. Данные, переведенные в электронную форму, приобретают новое качество, обеспечивая им более широкое распространение и эффективное использование. Современные информационные технологии пока не могут предоставить адекватный аппарат для оперирования с информацией и информационными ресурсами [1-3].

Сами по себе данные не представляют никакой информационной ценности без соответствующих описаний или моделей. Применение информационных технологий должно основываться на использовании конкретных моделей. А.А. Ляпунов [4] неоднократно отмечал: «нет модели – нет информации». Для возможности продуктивной работы нужны данные, превращенные в «информацию», представленную в виде «знаний», т.е. в форме «адекватного отражения действительности в сознании человека в виде представлений, понятий, суждений теорий» [5, 6].

Информационные потребности. В процессе научной, а особенно образовательной, деятельности много времени и сил отнимает работа с литературными источниками, разного рода материалами и документами, т.е. поиск необходимых документов, систематизация и классификация документов в соответствии с поставленной задачей.

Для удовлетворения информационных потребностей современных пользователей необходима поддержка сложных функций поиска и классификации информации, а также просмотр ресурсов по категориям (рубрикам) и словарям-классификаторам [3].

Современный студент, вооруженный компьютером, повседневно использующий возможности сети Интернет, не может быть удовлетворен традиционным режимом учебного процесса и обычными форматами учебных материалов, как-то учебники, книги или плоские текстовые файлы. Систематизация и классификация имеющихся информационных ресурсов в соответствии с потребностями пользователя является одной из важнейших задач поддержки как научной, так и образовательной деятельности.

Наиболее важной задачей является задача систематизации ресурсов, для решения которой необходимо четко определить состав логико-семантических категорий (фасетов) и ключевых терминов (понятий), покрывающих избранную достаточно узкую

предметную область, интересующую пользователя. Как правило предметная область ограничивается изучаемым учебным курсом или конкретной темой курса.

Стандартным подходом к систематизации информации является классификация документов с помощью таксономий. Таксономия – это предметная классификация, которая группирует термины в виде управляемого словаря (тезауруса) и упорядочивает их в виде иерархических структур. Основу классификации составляет выделение понятий (ключевых терминов), установление парадигматических отношений (например, типа родитель – потомок) между ними и сопоставление анализируемого документа выделенным понятиям.

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. По определению международной организации по стандартизации (ISO), тезаурус является словарем, управляемым языком индексации, формально организованным для того, чтобы установить явные априорные отношения между понятиями.

Индексация (индексирование) – это сортировка (систематизация) информации по каким-либо формальным принципам. Различают два вида индексирования: классификационное и координатное.

Тезаурусные (парадигматические) отношения (род-вид, часть-целое, комплекс-элемент, причина-следствие) налагаются на структуру таксономии, то есть идентифицируются основные таксономии предметной области.

Тезаурус – это словарь (идеографический или семантический словарь), в котором слова (лексические единицы, термины, понятия, дескрипторы) размещаются по их смысловой близости.

Тезаурусы в описании информации. С середины 1950 годов термин «тезаурус» прочно вошёл в профессиональную лексику специалистов в области информатики, причем определения тезауруса несколько варьировались в зависимости от класса задач, для решения которых он предназначался. В частности, применительно к задачам информационного поиска под тезаурусом обычно понимается так называемый нормативный тезаурус [7, с. 432] – словарь-справочник, содержащий все лексические единицы информационно-поискового языка – дескрипторы (вместе с ключевыми словами, которые в пределах данной информационно-поисковой системы считаются синонимами этих дескрипторов). Однако в 1990-х годах в информатике, наряду с термином «тезаурус», стал употребляться близкий по смыслу термин «онтология», при этом наиболее широко известно следующее определение Тома Грубера: «онтология – это явная спецификация концептуализации» (т.е. абстрактного представления предметной области) [23]. Применительно к информационным системам он обозначает «способ, который используется для описания некоторой области знаний..., в частности базовых понятий этой области, их свойств и связей между ними» [24]. В настоящее время, как отмечено в [22], под онтологией нередко стали понимать широкий спектр структур, представляющих знания о той или иной предметной области с разной степенью формализации [1, с. 30]:

1) словарь с определениями;
2) простая таксономия;
3) тезаурус (таксономия с терминами);
4) модель с произвольным набором отношений;
5) таксономия и произвольный набор отношений;
6) полностью аксиоматизированная теория.

В работах многих авторов термин «онтология» начал употребляться вместо термина «тезаурус» (что, в общем, неудивительно, ибо определения онтологии в той или иной степени сходны с определением тезауруса, а первоначальное значение термина «онтология» – «учение о бытии», звучит куда более многообещающе, чем заурядный «тезаурус» – «сокровище» или «запас»).

Тезаурус является инструментом концептуального (понятийного) описания отдельных предметных областей. В отличие от толкового словаря, он определяет понятия не только с помощью описания, данного в словарной статье, но и посредством соотнесения понятия с другими понятиями и их группами, благодаря чему, как отмечено Шрейдером Ю. А. [9, 10], может использоваться как система знаний, отраженных языком (словами).

Для описания какой-либо предметной области всегда используется определенный набор ключевых терминов, каждый из которых обозначает или описывает какое-либо понятие (концепцию) из данной предметной области. Совокупность терминов, описывающих данную предметную область, с указанием семантических отношений (связей) между ними называется тезаурусом. Такие отношения в тезаурусе всегда указывают на наличие смысловой (семантической) связи между терминами.

Основным отношением (связью) между терминами в тезаурусе является связь между более широкими (более выразительными) и более узкими (более специализированными) понятиями. Часто выделяют два подвида этого отношения: термин обозначает понятие, являющееся частью понятия, обозначаемого другим термином (например, «наука» и «математика», «математика» и «теория чисел»); термин обозначает элемент класса, обоначае-мого другим термином («горные районы» и «Кавказ»).

Существуют и другие связи между терминами. Например, одно понятие может быть обозначено несколькими терминами – синонимами. Некоторые термины могут быть антонимами для других. Часто среди терминов, относящихся к одному понятию, выделяют единственный (для каждого языка тезауруса) наиболее предпочтительный (самый подходящий) термин, который наилучшим образом характеризует, или обозначает данное понятие. Остальные термины являются менее предпочтительными (менее подходящими).

Помимо вышеописанных, между терминами могут быть и другие отношения, если понятия, обозначаемые этими терминами, как-либо связаны между собой по смыслу (ассоциативные связи), за исключением описанных выше иерархических связей.

Термины могут обладать следующими атрибутами (основные) [3]:

ID – Identifier. Уникальный идентификатор термина.

SN – Scope Note. Комментарий к термину. Например, представляет вербальное пояснение термина, или правила его использования.

TT – Top Term. Признак, выделяющий термины на самом верхнем уровне иерархии (термины наиболее общих понятий в данной иерархии понятий).

HN – History Note. История модификации связей и атрибутов данного термина.

Термины могут иметь следующие отношения (основные):

USE – Связывает термин с наиболее предпочтительным (на том же языке) термином для данного понятия.

UF – Used For. Обращение связи USE. Связывает наиболее подходящий термин с синонимами и квазисинонимами (менее подходящими терминами).

BT – Broader Term. Связь термина с термином более общего (родительский термин) понятия.

NT – Narrower Term. Связь термина с термином более узкого (дочерний термин) понятия. Обращение связи BT.

RT – Related Term. Ассоциативная связь. Связывает семантически связанные между собою термины, не находящиеся при этом в одной иерархии, и не являющиеся синонимами или квазисинонимами.

LE – лексический эквивалент термина (на другом языке).

Истоки тезауруса в информационном мире. История возникновения тезаурусов восходит к великим мыслителям древности и, прежде всего, к Аристотелю. Ему принадлежат слова, возраст которых 2,5 тыс. лет: «Из слов, высказываемых без какой-либо связи, каждое означает или сущность, или качество, или количество, или отношение, или обладание, или действие, или страдание» [11].

Одной из наиболее древних попыток идеографических классификаций является труд «Аттический глоссарий» Аристофана Византийского (в 195 – 180 гг. до н.э. – директор Александрийской библиотеки), упоминаемого также как составителя одного из первых толковых словарей греческого языка.

В I веке н. э. появляется «Словарь синонимов» грекоязычного ученого Филона Библского, который называют одним из первых тезаурусов. Филон Библский (Геренний Филон) – финикийский историк, грамматик, родом из города Библа. Он составил на греческом языке труды по истории и мифологии Финикии «О городах и о том, что замечательного в них было» в 30 книгах, «О приобретении и отборе книг» в 12 книгах, «Словарь синонимов», «Об иудеях».

Во II веке н.э. появляется капитальный труд на материале греческого языка, составленный лексикографом и софистом Юлием Поллуксом (Ю́лий Полиде́вк), уроженцем египетского города Навкратис. Он пользовался покровительством римских императоров Марка Аврелия и его сына Коммода и, преподавая риторику в Афинах, написал несколько сочинений, из которых до нас дошел лишь его словарь «Ономастикон» («Искусство наименования»), причём не в оригинальном варианте, а в более позднем переложении [12].

Словарь Ономастикон состоит из 10 книг. Каждая из них начинается приветственным письмом к императору Коммоду. Книги по существу являются отдельными трактатами и содержат в себе наиболее важные слова, относящиеся к той или иной теме. Так, в первой книге говорится о богах и царях, во второй – о людях, их жизни и фи-

зиологическом строении, в третьей – о родстве и гражданских отношениях и т. д. Слова, помещенные в словарь, сопровождаются краткими толкованиями. В новое время словарь был впервые опубликован в 1502 г. в Венеции. Словарь составлен по предметному принципу, с включениями элементов риторики, содержит краткие толкования слов, множество цитат более ранних авторов: Аристофана Византийского, Эратосфена, Памфила и других, сведения из разных областей жизни, а также 52 выражения, служащие для восхваления правителей, и 33 выражения для ругательств в адрес сборщиков податей [12].

Между II и III веками н. э. появился санскритский словарь «Амаракоша» (сокровище Амара), другое название «Намалинганушасана», – наставление о существительных и грамматических родах, который содержал около 10 тыс. слов и состоял из трех книг, каждая из которых делилась на главы, главы – на секции. Так, первая книга была посвящена небу, богам и всему тому, что с ними связано. В ней имелись секции «времена года», «небесный свод» и т.п. Вторая книга содержала слова, относящиеся к земле, растительному и животному миру и человеку. Автором словаря является древнеиндийский поэт, грамматик и лексикограф Амара Сина (по другим источникам Амара Дэва). Словарь составлен в строгом соответствии с уровнем развития науки и господствующими представлениями в обществе, что позволяет нам судить о понимании и объяснении мира, об устройстве человеческого общества в те далекие времена. Словарь Амаракоша разительно отличается от алфавитных словарей. Так, если в алфавитных словарях порядок записей (терминов) регулируется алфавитом, являющимся в значительной мере нейтральным инструментом, то при построении идеографического словаря решающее значение приобретает мировоззрение самого лексикографа. Даже тот факт, что автор словаря был буддистом, нашел отражение в соответствующем разделе словаря [12].

Термин «тезаурус» – «сокровище», аналог санскритского термина «коша», был применен впервые в XIII веке в заголовке труда флорентийского поэта и ученого Брунето Латини (Brunetto Latini) – систематизированной энциклопедии, названной им «Книга о сокровище» [13]. Она была издана во Франции под названием «Livres dou trésor» и содержала обзор знаний своего времени о Боге, природе, об истории древнего и нового времени, об искусстве, а также давала наставления для управления домом, государством и т.д. Это вполне соответствовало семантике употребленного греческого слова «Thesauros» (греч. $\theta\eta\sigma\alpha\nu\rho\varsigma$, т.е. «сокровище», «богатство», «запас»).

Следует подчеркнуть, что первые тезаурусы составлялись без всякой связи с особенностями информационной деятельности, они были органически связаны с фундаментальными проблемами познания, отображая представление о мире в целом и закономерностях его постижения средствами естественного языка.

В дальнейшем, при проникновении идей тезауруса в автоматизированные информационно-поисковые системы (ИПС), тезаурус стал рассматриваться как словарь для построения поисковых образов документов и запросов, качество которых существенно влияло на качество поиска информации.

В первой половине XIX века коллежский советник Семён Николаевич Корсаков поставил задачу усиления возможностей разума посредством разработки научных ме-

тодов и устройств. В 1832 году он опубликовал описание пяти изобретённых им механических устройств, так называемых «интеллектуальных машин», для частичной механизации умственной деятельности в задачах поиска, сравнения и классификации. В конструкции своих машин Корсаков впервые в истории информатики применил перфорированные карты [14]. В работах Корсакова содержится целая плеяда новых для того времени идей, как-то: многокритериальный поиск с учетом относительной степени важности различных критериев (весовых коэффициентов), способ обработки больших массивов данных, предтеча современных экспертных систем, попытка определить понятие алгоритма. Он предложил общий принцип поиска или сравнения идей (записей, фактов) на основе их деталей (признаков). Таким образом, С.Н. Корсаков, выражаясь современным языком, определил информационную запись набором ее признаков – координатное индексирование. Словарь Корсакова представлял перфорированные таблицы, где каждый столбец определял некоторую идею, а в строках перфорировались признаки этой идеи. Перфорация отверстий обеспечивала возможность механического поиска и сравнения идей на основе их признаков. Опубликование Корсаковым работ на французском языке, который являлся общепризнанным международным языком того времени, закрепило приоритет за русским изобретателем.

Один из первых в истории и наиболее известных на сегодня тезаурусов (идеографических словарей) составлен британским лексикографом Питером Марком Роже и опубликован в 1852 г. (тезаурус Роже – Roget's Thesaurus). Его оригинальное название –Thesaurus of English Words and Phrases. Непосредственными предшественниками словаря П. М. Роже были работы Д. Далгарно, Д. Уилкинса и санскритский словарь «Амаракоша», переведенный на английский язык в 1808 г. [12].

Роже в полной мере использовал опыт своих предшественников. Он пишет: «Принцип, которым я руководствовался, классифицируя слова, является тем же самым, который используется при классификации особей в различных областях естественной истории. Поэтому разделы, выделенные мной, соответствуют естественным семьям ботаники и зоологии, а ряды слов сцементированы теми же отношениями, которые объединяют естественные ряды растений и животных».

В предисловии к своему словарю он пишет: «...Какой бы живостью ни обладало наше воображение, как бы ни переполняли нас чувства, мы часто попадаем в такое положение, когда нам не хватает слов, чтобы точно выразить свою мысль. Единственно необходимое слово зачастую бежит нашей памяти, и мы вынуждены обходиться словами слишком сильными или слишком слабыми, слишком общими или излишне конкретными. Помощь, которую оказывает этот словарь, состоит в предоставлении богатейшего набора слов и выражений, исчерпывающих все оттенки и нюансы каждой общей идеи».

Дальнейшее развитие идеи смысловой классификации лексики связано с проблемой так называемого всемирного философского языка. В основании этого языка должна лежать логическая классификация всего, что может быть предметом человеческой мысли.

Проблема формализации работы разума восходит к Аристотелю, а вопросы создания языка разума волновали Декарта. Значительный вклад в разработку этого во-

проса внес испанский философ, поэт, богослов и миссионер Раймонд Луллий (Raymundus Lullius) – один из наиболее ярких и оригинальных мыслителей европейского Средневековья [15, 16]. Его книга «Великое и окончательное искусство» («Ars Magna et Ultima») заметно повлияла как на современников, так и наше время. В этой книге он описал возможную конструкцию логической машины, предложив оригинальную систему, состоящую из семи концентрических кругов. Каждый круг содержал обозначение группы понятий. Так, на одном из них приводятся девять субстанций: бог, ангел, небо, человек, воображаемое, чувственное, растительное, стихийное, инструментальное. Другой круг служил для указания на девять абсолютных предикатов: святость, величина, длительность, могущество, знание, стремление, добродетель, истина, слава; третий содержал девять относительных предикатов: святой, великий, длительный, могучий, добродетельный и т. п.

Кроме того, элементами системы испанского философа были понятия, отражающие различные отношения между предметами (различие, тождество, противоречие, начало, середина, конец, больше, равно, меньше) и список вопросов (что, чего, почему, как велико, какого качества и др.). При вращении кругов относительно друг друга возникали различные комбинации понятий, что обеспечивало возможность, по мнению Р. Луллия, находить истины и перечислять все предметы мысли.

Одной из важных заслуг Р. Луллия у современников считалось подробное описание видов и сущностей хаоса – первоэлементов, которыми обозначались общие понятия или основные категории всего существующего, из которых создается мир, и отношений между ними, которые напоминают современные RDF-диаграммы.

«Искусство памяти» Луллия (логическая машина) базируется на системе, посредством которой возможно формулировать множество теологических вопросов и ответов. В нем используются символы, алфавит, обозначающий различные вещи, и потому кажется сходным с Аристотелевской символической логикой. Сами буквы не наделены каким-либо значением, однако они выступают в роли определенных понятий, взятых из теологии и логики, которые и обозначают. С другой стороны, «Искусство» обладает своего рода гибкостью, которая не была присуща Аристотелевской логике: та логика является словесной, она требует от оратора «вручную» подготавливать каждый из аргументов [11].

Кроме того, логика Аристотеля является синтаксической, а не семантической; то есть сказать, что «Все собаки - кошки; я - собака, следовательно, я - кошка», - возможно, но хотя утверждение является логичным, само заявление является ложным.

Многие математики рассматривают устройство Луллия как первую в истории попытку создать машину, способную выполнять ряд логических операций, осуществляемых человеческим мозгом. Луллий хотел свести процесс исследования реальности к механическому комбинированию небольшого количества исходных аксиом, отражающих отдельные фрагменты этой реальности. Он считал, что в этих аксиомах неявно содержатся все истины науки, и что их можно извлечь путём чистого мышления, без обращения к опыту. Прежде всего, Луллий был миссионером, но, в отличие от инквизиторов, он, францисканец, хотел нести веру не крестом и мечом, а логикой и стремился быть доказательным в своей позиции.

Он хотел говорить с иноверцами на их языке, используя для этого, как бы сегодня сказали, формальную логику. Ему были знакомы алгебраические труды Аль Хорезми и само понятие «алгоритм», а во время путешествий по Северной Африке он видел гадальные устройства, состоящие из концентрических вращающихся дисков. Сочетая алгоритмическую основу с дисковой механикой, Луллий намеревался создать инструмент для получения логических доказательств.

Логическая машина Луллия была описана Джонатаном Свифтом, англо-ирландский писателем-сатириком, в знаменитой фантастической тетралогии «Путешествия Гулливера». В третьей части «Путешествий», где описан визит Гулливера в Великую академию, находящуюся в столице Лапутии, автор высмеивает некоторых не слишком прагматичных ученых. Переходя из комнаты в комнату, Гулливер в одной из них обнаруживает своеобразную конструкцию и выясняет, что огромная, площадью 20 квадратных метров рама представляет собой прибор для открытия отвлеченных истин. На раме располагаются таблички со словами, которые можно произвольно сочетать с помощью встроенного «генератора случайных чисел». Совершив очередную перетасовку, вовлеченные в эксперимент ученые пытаются найти в беспорядочном наборе слов и знаков осмысленные фразы и таким образом создать полный обзор всех наук и искусств. Эту машину Свифт назвал «компьютером». Свифт использовал это название вслед за Лейбницем, который считал, что научный спор можно решить с помощью бумаги, пера и вычислений (Gentlemen, let us compute!)

Надо иметь в виду, что до недавнего времени под словом compute понимали доказательство чего-то математическими средствами, а не просто как вычисление.

Особое влияние идеи Луллия оказали на формирование философских и научных воззрений Г. В. Лейбница. Он писал о возможности создания всеобщего философского языка, в основании которого лежала бы некоторая система понятий, и даже придумал название для этого языка (Specieuse gcnerale).

Так, Лейбниц заявлял, что его цель – не просто получение новых результатов (например, в математике), а выработка общего формального метода, позволяющего находить таковые. В основе, по Лейбницу, должен был лежать «всеобщий алфавит» человеческих знаний. Но, в отличие от Луллия, Лейбниц считал необходимым сверх того создать для них целесообразную систему обозначений – универсальную характеристику, а также выработать систему правил для получения сложных понятий из простейших. В позднейших философских сочинениях Лейбниц даже заявляет, что Бог творит путем исчисления возможных миров и выбора наилучшего, т. е. фактически по методике Луллия.

В начале XVII века с обоснованием идеи создания всемирного философского языка выступает Р. Декарт, считающий, что в основе такого языка должна лежать классификация всех объектов человеческого мышления.

В 1661 году Джордж Дальгарно издал трактат, посвящённый философскому (или универсальному) языку. По существу, Д. Дальгарно разработал словарь, основанный на логической классификации понятий, представляющий собой систему философского языка, построенного с помощью условных обозначений категорий и соотношений посредством букв [17].

Всего Д. Дальгарно выделяет 17 классов, закрепляя за каждым определенную согласную букву (например, n – конкретные предметы, k – отношения и т. п.). Каждый класс разделяется на подклассы, которые в свою очередь обозначаются второй буквой, гласной (например, ka – служебные отношения, ki – партийные отношения, ku – отношения вражды). В результате подобного разбиения получаются такие, например, слова: Nyka– слон, Nykn – лошадь, Nyke – осел. Сразу же после выхода в свет книги Дальгарно она подверглась критике, что неправильно классифицировала научное знание. Выдающийся учёный того времени, епископ Джон Уилкинс, попытался исправить недостаток подхода Дальгарно, опираясь на гораздо больший объём сведений.

Джон Уилкинс в 1668 году издает трактат, в котором он предложил универсальный язык и десятичную систему мер, которая впоследствии стала основой метрической системы. Все понятия, охватываемые языком, Д. Уилкинс делит на шесть типов: трансцендентальные понятия, субстанции, количества, качества, движения, отношения. Затем эти шесть типов в свою очередь подразделяются на 40 классов, каждый из которых обозначается определенным слогом: Ba, Ba, Be, Bi; Da, Da, De, Di; Ga, Ga, Ge, Gi; и т. п. Присоединение к этим слогам согласных b, d, g, p, t, c, z, s, n приводит к получению определенного числа родов. Указанные согласные, будучи показателями родов, имеют строго определенные значения: b – первый род, d – второй род, g – третий род и т. п. [18]

Идеографическая классификация. Исходя из общих методологических принципов классификации понятий можно говорить о трех типах идеографических словарей [19 -21]:

Идеографический словарь-тезаурус основан на логической рубрикации всего понятийного содержания лексики. Главной задачей является идентификация и последующая рациональная классификация понятийных групп, реально представленных в лексике языка, характер и количество лексико-семантических групп определяется смысловой емкостью языка.

Аналогический словарь-тезаурус основан на выделении тематических лексико-семантических групп, которые располагаются в порядке алфавитного следования тематических доминант (слов-центров).

Тематический словарь-тезаурус основан на выделении тем, характеризующих выбранную предметную область.

Основные требования, которые предъявляются к классификационной системе, сводятся к следующему [12]:

- классифицироваться должны не слова определенного языка, а понятия, что обеспечивает универсальность системы;
- классифицируются понятия исходные, лежащие в основе языка;
- направляющим принципом классификации является осознание системы понятий как определенным образом организованного единства, расчленение которого должно вестись в естественной последовательности.

Тезаурусы в информационном поиске. В информационной системе тезаурус яв-

ляется не только самостоятельным информационным ресурсом, но и инструментом для классификации или индексации ресурсов.

Исторически тезаурусы создавались для ручного индексирования документов и при их создании не принимались во внимание вопросы, связанные с автоматической индексацией. Трудность построения тезауруса, соответствующего всему тематическому многообразию индексируемой информации, является основной причиной его непопулярности в современных информационных системах. Но эффективность информационно-поисковых систем для поддержки научно-образовательной деятельности напрямую зависит от использования специализированных тезаурусов.

Пользователь информационной системы должен иметь возможность [1, 22, 26]:

- осуществлять просмотр тезауруса;
- осуществлять поиск ресурсов по ассоциированным с ними терминам или понятиям.

Поиск ресурсов может вестись двумя способами:

- поиск, по ключевым словам, с использованием тезауруса;
- навигация по тезаурусу.

При поиске ресурсов, по ключевым словам, поисковая система может, используя тезаурус, расширять результаты поиска, выдавая пользователю не только ресурсы, соответствующие введенным пользователем ключевым словам, но и ресурсы, соответствующие связанными с ними терминами, или терминами, обозначающими более узкие понятия относительно исходных терминов.

Например, если пользователь ищет ресурсы, соответствующие термину «тунель», в результатах поиска необходимо выдать также все ресурсы, соответствующие термину «тоннель», поскольку оба они являются разными вариантами написания одного и того же слова. Или если ищутся ресурсы, соответствующие понятию «проблемы управления», имеет смысл включить в результаты поиска также ресурсы, соответствующие рубрике «задачи оптимального управления».

Система поиска может также, используя тезаурус, подсказать пользователю, по каким еще словам ему стоит осуществить поиск (например, квазисинонимы, связанные термины, более широкие термины, и т.д.). Оба этих варианта использования тезауруса широко применяются, например, в поисковых машинах. Интерфейс просмотра тезауруса должен:

- показывать все атрибуты данного термина или понятия;
- показывать, с какими терминами и понятиями связан данный термин или понятие.

Достаточно наглядно показывать пользователю место термина или понятия в иерархии понятий тезауруса.

Однако существует ряд тезаурусов, основной задачей которых является не индексация ресурсов, а их классификация. В этом случае основными объектами тезаурусов (классификаторов) выступают не термины, а понятия (рубрики), и, часто, идентифицирующие их уникальные идентификаторы (коды классификации). Отношения в таком тезаурусе – не семантические связи между терминами, а характеризующие логику описываемой предметной области отношения между понятиями (рубриками).

Структура классификатора соответствует структуре обычного тезауруса, посколь-

ку связи между его рубриками по смыслу те же, что и между терминами тезауруса, и классификатор является его частным случаем. Однако при классификации в соответствие ресурсам ставятся не термины, а обозначаемые ими понятия. Потому в схеме данных информационной системы понятия тезауруса должны быть выделены в самостоятельные объекты.

Заключение. Рассмотрена история развития и возникновения тезаурусов. Дан исторический обзор основных подходов и принципов работы великих мыслителей прошлого. Рассмотрены вопросы использования тезауруса, идеографических, управляемых словарей в информационном поиске. Определены интересные перспективы дальнейшего использования многоязычного тезауруса для информационной системы поддержки в научно-образовательной деятельности, способной в автоматизированном режиме извлекать поисковые признаки из цифровых документов достаточно произвольной структуры для систематизации и классификации документов.

ЛИТЕРАТУРА

[1] *Шокин Ю.И., Федотов А.М., Барахнин В.Б.* Проблемы поиска информации. Новосибирск: Наука. 2010. 196 с.

[2] *Федотов А. М., Барахнин В. Б.* Проблемы поиска информации: история и технологии // Вестник НГУ. Серия: Информационные технологии. 2009. Т. 7, вып. 2. С. 3-17.

[3] *Федотов А.М., Идрисова И.А., Самбетбаева М.А., Федотова О.А.* Использование тезауруса в научно-образовательной информационной системе // Вестник НГУ. Серия: Информационные технологии. 2015. Т. 13, вып. 2. С. 86-102.

[4] *Ляпунов А.А.* О соотношении понятий материя, энергия и информация // В кн.: Ляпунов А. А. Проблемы теоретической и прикладной кибернетики. Новосибирск: Наука, 1980. С. 320-323.

[5] *Федотов А.М., Барахнин В.Б., Жижимов О.Л., Федотова О.А.* Модель информационной системы для поддержки научно-педагогической деятельности // Вестник НГУ. Серия: Информационные технологии. 2014. Т. 12, вып. 1. С. 89-101.

[6] *Федотов А.М., Абделиева М.Н., Байдавлетов А.Т., Бапанов А.А., Самбетбаева М.А., Федотова О.А.* Концептуальная модель научно-образовательной информационной системы // Вестник НГУ. Серия: Информационные технологии. 2015. Т.13, вып. 3. С. 52-67.

[7] *Михайлов А. И., Черный А. И., Гиляревский Р. С.* Основы информатики. М.: Наука, 1968.

[8] *Шрейдер Ю.А.* Равенство, сходство, порядок. М.: Наука, 1971.

[9] *Шрейдер Ю. А.* О количественных характеристиках семантической информации // НТИ: Сер. 2. 1963, № 10. С. 35–39.

[10] *Шрейдер Ю.А.* Об одной модели семантической информации // В сб.: Проблемы кибернетики. Вып. 13. М.: Наука, 1965. С. 233-240.

[11] *Аристотель.* Аналитика: Первая и вторая: Пер. с греч. / Аристотель. – [б. м.] Госполитиздат, 1952. 438 с.: 7.75.

[12] *Морковкин В. В.* Идеографические словари. М.: Из-во МГУ, 1970. 72 с.

[13] *Брунетто Латини* // Энциклопедический словарь Брокгауза и Ефрона: в 86 т. (82 т. и 4 доп.). СПб., 1890-1907.

[14] *Karsakof S.* Apercu d`un procédé nouveau d`investigation au moyen de machines à comparer les idées. St. Petersbourg, 1832.

[15] *Бирюков Б.В., Тростников В.Н.* Жар холодных чисел и пафос бесстрастной логики. Формализация мышления от античных времен до эпохи кибернетики. М.: Знание, 1977, 2004.

[16] *Renan E., Lulle R.* Histoire littéraire de la France, par les membres de l'Institut, Vol. XXIX. – Париж, 1885.

[17] *Борхес Х.Л.* Аналитический язык Джона Уилкинса // Собрание сочинений. Т. 2. Амфора, 2005. С. 416-420.

[18] *Дрезен Э.* За всеобщим языком. Три века исканий. М.: Едиториал УРСС, 2004.

[19] *Аджиев А.С., Бездушный А.Н., Серебряков В.А.* О реализации веб-системы математической информации [Электронный ресурс] / Российский научный электронный журнал Электронные библиотеки, 2004. Т. 7, вып. 1.

[20] *Михайлов А.И., Черный А.И., Гиляревский Р.С.* Научные коммуникации и информатика. М: Наука, 1976.

[21] *Арский Ю.М., Гиляревский Р.С., Туров И.С., Черный А.И.* Инфосфера: Информационные структуры, системы и процессы в науке и обществе. М.: ВИНИТИ, 1996.

[22] *Лукашевич Н. В.* Тезаурусы в задачах информационного поиска. М.: Издательство МГУ, 2011. С. 512.

[23] *Gruber T.* A translation Approach to Portable Ontology Specifications // Knowledge Acquisition Journal. 1993. Vol. 5, No. 2. P. 199–220.

[24] *Овдей О. М., Проскудина Г. Ю.* Обзор инструментов инженерии онтологий // Труды Шестой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2004). Пущино, 2004. С. 59–68.

[25] *Нариньяни А.С.* Кентавр по имени ТЕОН: Тезаурус + Онтология // Труды международного семинара Диалог-2001 по компьютерной лингвистике и ее приложениям. Аксаково, 2001. Т. 1. С. 184-188.

[26] *Нариньяни А.С.* ТЕОН-2: от Тезауруса к Онтологии и обратно // Труды международного семинара Диалог-2002 по компьютерной лингвистике и ее приложениям. Протвино, 2002. Т. 1. С. 307-313.

[27] *Федотов А. М., Барахнин В. Б.* К вопросу о поиске документов «по аналогии» // Вестник НГУ. Серия: Информационные технологии. 2009. – Том 7, вып. 4. С. 3-14.