DOI 10.24069/konf-23-26-04-2019.04

Разработка национального стандарта на сопоставление информационно-поисковых языков

В. Н. Белоозеров, О. А. Антошкова
Всероссийский институт научной и технической информации РАН, г. Москва, Россия

Резюме: Изложено обоснование разработки и содержание проекта национального стандарта России на сопоставление и взаимодействие словарей информационно-поисковых языков. Стандарт разрабатывается как модифицированный аналог ISO 25964-2:2014. Проект описывает широкий спектр словарей, взаимодействие которых можно использовать для поиска информации в среде разнородных информационных ресурсов. Рассмотрены следующие типы словарей: информационно-поисковые тезаурусы, классификационные системы различного назначения, словари предметных рубрик, онтологии, терминологические словари, словари синонимов и авторитетные файлы имен. Определены смысловые соответствия, которые целесообразно устанавливать между значащими элементами словарей при их сопоставлении для взаимодействия. Установлены обозначения соответствий. Даны рекомендации по сопоставлению с тезаурусами каждого вида рассмотренных словарей.

Ключевые слова: информационно-поисковый язык; предметное индексирование; тезаурус; классификация; словарь предметных рубрик; терминологический словарь; словарь синонимов; онтология; авторитетный файл; сопоставление словарей

Для цитирования: Белоозеров В. Н., Антошкова О. А. Разработка национального стандарта на сопоставление информационно-поисковых языков // Научное издание международного уровня – 2019: стратегия и тактика управления и развития: материалы 8-й Международ. науч.-практ. конф., Москва, 23–26 апреля 2019 г. Екатеринбург: Изд-во Урал. ун-та, 2019. С. 35–43. DOI: 10.24069/konf-23-26-04-2019.04.

Development of a national standard for comparison of information retrieval languages

V. N. Beloozerov, O. A. Antoshkova
All-Russia Institute of Scientific and Technical Information of Russian Academy of Science, Moscow, Russia

Abstract: The rationale for the development and the content of the draft national standard of Russia on the matching and interaction of the vocabularies of information retrieval languages are stated. The standard is developed as a modified analogue of ISO 25964-2:2014. The draft describes a wide range of vocabularies, the interaction of which can be used to search in the environment of heterogeneous information resources. The following types of vocabularies are considered: information-retrieval thesauri, classification schemes for various purposes, subject heading schemes, ontologies, terminologies, vocabularies of synonym rings, and name authority lists. Semantic correspondences, which are expedient to establish between significant elements of the vocabularies at their comparison for interaction, are defined. A set of symbols for the correspondence of lexical units is established. Recommendations for comparison with thesauruses of each type of the considered layers are given.

Keywords: information languages; thesaurus; classification; subject headings dictionaries; terminological dictionaries; dictionaries of synonyms; ontologies; indexing dictionaries; taxonomy; dictionaries matching; information retrieval

For citation: Beloozerov V. N., Antoshkova O. A. Development of a national standard for comparison of information retrieval languages. In: *World-Class Scientific Publication – 2019: Strategy and Tactics of Management and Development: Proc. 8th Int. Sci. & Pract. Conf., Moscow, April 23–26, 2019.* Ekaterinburg: Ural University Press, 2019, pp. 35–43. DOI: 10.24069/konf-23-26-04-2019.04.

Введение

Задача установления смыслового соответствия классов различных классификационных систем является традиционной для сферы научной и технической информации. Так, Государственный рубрикатор научно-технической информации (ГРНТИ), разработанный в 70-х гг. пошлого века, непосредственно в своих таблицах имеет сопоставления с классами Универсальной десятичной классификации (УДК) и позициями Номенклатуры специальностей научных работников Высшей аттестационной комиссии (ВАК). Также традиционным является мнение о невозможности адекватного установления эквивалентности между классами различных систем классификации, что послужило основанием для свертывания этих работ на рубеже XX и XXI вв. Однако развитие Интернета и концепция семантического веба открывают возможность реального информационного поиска и навигации по сети разнородных ресурсов, когда необходимость переходить от одной системы организации знаний в одном ресурсе к другой системе в другом ресурсе переводят проблему соответствия способов описания содержания ресурсов из теоретической плоскости в практическую, когда при невозможности точного установления эквивалентности рубрик информации следует находить возможность неточного перехода с одного ресурса на другой с указанием степени неточности сохранения тематики контента. При этом речь идет о сопоставлении не только классификационных систем, но также и других средств индексирования и поиска информации, использованных в среде доступных ресурсов. Практическая потребность сопоставления систем организации знаний в различных информационных ресурсах привела к определенному оживлению этих работ. В нашей стране в этом направлении ведутся исследования в ВИНИТИ РАН, БЕН РАН, РГБ и в ряде других организаций. На международном уровне эти работы поддержаны изданием международного стандарта ИСО 25964-2:2014 [1], в котором даны подробные рекомендации по установлению соответствий между словарями различного типа, которые используются как источник лексических единиц для тематического индексирования документов и/или запросов. В стандарте основное внимание уделено информационно-поисковым тезаурусам и их взаимодействию с другими информационно-поисковыми языками, но вполне содержательная информация дана и для взаимодействия классификационных систем (библиографических, документоведческих и рубрикаций баз данных), терминологических словарей, онтологий и словарей синонимов с тезаурусами и друг с другом.

Предполагая, что по мере развития практики поиска в глобальных сетях разнородных информационных ресурсов будет расширяться также практика и потребность взаимодействия словарей тематического индексирования, технический комитет Росстандарта ТК 191 «Научно-техническая информация, библиотечное и издательское дело» поставил в план национальной стандартизации разработку российского

стандарта путем адаптации международного стандарта ИСО 25964-2:2014 к условиям отечественной информатики. ВИНИТИ РАН разработал соответствующий проект стандарта, который обсуждается ниже.

Содержание стандарта

Российский ГОСТ разрабатывается как гармонизированный с международной стандартизацией и должен содержать все разделы ИСО 25964-2:2013 [1] за исключением не имеющих нормативной силы предисловия и приложения. Однако гармонизацию предполагается осуществить путем принятия модифицированного, а не идентичного международному национального стандарта, в который могут быть внесены точно обозначенные изменения положений исходного международного стандарта.

Основной текст стандарта можно представить состоящим из двух разделов. Главы с 1 по 16 касаются принципов и практических возможностей сопоставления, которые применимы к большинству типов словарей, но ориентированы главным образом на тезаурусы, структура и содержание которых определены первой частью международного стандарта ИСО 25964-1:2011 [2] или идентичным российским стандартом ГОСТ Р 7.0.91 [3]. Главы с 17 по 24 посвящены отдельным типам словарей. Рассмотрены словари, которые обычно используются для классификации или индексирования информационных ресурсов, а именно системы классификации (в том числе используемым для управления документацией), таксономии, словари предметных рубрик и авторитетные файлы имен. Также рассмотрены терминологические словари, онтологии и словари синонимов, несмотря на отличие в их назначении. В каждой главе дается краткое описание ключевых характеристик словарей в сравнении их смысловых единиц, чтобы обеспечить понимание рекомендаций по сопоставлению.

Особого внимания заслуживает глава ***3 Термины и определения***, где предложена целая терминосистема 89 понятий, описывающая разные виды словарей с точки зрения сопоставления семантики словарных единиц. Определены, в частности, понятия словарей, подлежащих сопоставлению: ***классификационная система; контрольный словарь; тезаурус; авторитетный список имён; онтология; структурированный словарь; словарь предметных рубрик; таксономия*** (*как вид словаря*). Однако здесь отсутствуют определения таких понятий как ***терминологический словарь и словарь синонимов,*** которые приходится вводить и разъяснять далее в главах, посвященных процедуре сопоставления. Определены виды словарных единиц: ***термин, предмет (тема), класс, наименование класса, классификационный код, категория*** *(в таксономии)*, ***категориальная метка, предметная рубрика, индексный термин, поисковый термин, входной термин, предпочтительный термин*** *(дескриптор)*, ***непредпочтительный термин*** *(аскриптор).* Однако в определениях этих понятий отсутствует четкая соотнесенность видов словарных единиц с видами словарей и с ролью в информационном процессе. Определены виды смысловых отношений словарных единиц: ***ассоциативные отношения; иерархические отношения; вышестоящий термин; нижестоящий термин; эквивалентное соответствие; отношение эквивалентности; соответствие «один-к-одному»; соответствие «один-ко-многим».*** Здесь четко разделены внутрисловарные смысловые

связи, определяемые как «отношения», и межсловарные связи, определяемые как «соответствия». Другие термины, которым даны определения в этой главе стандарта, фрагментарно описывают внутреннюю структуру словарей и некоторые детали процессов индексирования и информационного поиска.

В тексте исходного международного стандарта определения терминов расположены в алфавитном порядке, без какого-либо смыслового структурирования. Это затрудняет оценку системности принятой терминологии. После перевода на русский язык этот недостаток оригинала усугубился тем, что и алфавитный принцип расположения терминов нарушется. Целевая аудитория нового стандарта должна высказаться по поводу того, следует ли оставить термины в национальном стандарте под теми же номерами и в том же порядке, как они представлены в международном стандарте, или же целесообразно их структурировать по-новому в логической последовательности.

Глава 4 определяет обозначения типов соответствия единиц разных словарей, которые можно представить в структурированном виде (см. табл. 1). Однако точное описание формата утверждений о соответствии (формулы соответствия) отсутствует. Даются только разрозненные примеры записей.

Таблица 1. **Обозначения соответствий**

Символ соответствия		*Значение*	*Примечание*
русский	*английский*		
ЭК		Эквивалентность	Значения терминов совпадают
=	=	Точное соответствие	–
~	~	Неточное соответствие	–
ШС	BM	Широкое соответствие	Термин слева шире термина справа
ШСР	BMG	Соответствие родовому термину (широкое соответствие)	Термин справа есть родовое понятие для термина слева
ШСМ	BMI	Соответствие термину множества (широкое соответствие)	Термин справа есть множество, содержащее денотат термина слева
ШСЦ	BMP	Соответствие термину целого (широкое соответствие)	Термин справа есть целое, часть которого – денотат термина слева
УС	NM	Узкое соответствие	Термин слева уже термина справа
УСВ	NMG	Соответствие видовому термину (узкое соответствие)	Термин справа есть видовое понятие для термина слева
УСЭ	NMI	Соответствие элементу множества (узкое соответствие)	Термин справа есть элемент множества, обозначенного термином слева
УСЧ	NMP	Соответствие термину части (узкое соответствие)	Термин справа есть часть того, что обозначено термином слева
АС	AM	Ассоциативное соответствие	Термины обозначают ассоциативные (смежные) понятия
|	|	Объединение значений	Соответствие установлено для объединения значений терминов (экстенсионалов понятий)
+	+	Пересечение значений	Соответствие установлено для пересечения значений терминов (экстенсионалов понятий), или, что то же, для объединения интенсионалов понятий

В главе 5 сделана попытка определить, какой элемент в каждом типе словарей должен представлять понятие в утверждениях о соответствии, а также указано, что основная цель сопоставления словарей – обеспечить возможность преобразования поисковых образов документов и поисковых предписаний при поиске информации в разнородных источниках.

Глава 6 содержит обзор возможных структур установления соответствий в совокупностях нескольких сопоставляемых словарей. Основных структур три: (1) взаимно однозначное соответствие всех словарей, (2) установление соответствий для каждой пары словарей и (3) «звездная» структура – соответствия каждого словаря устанавливаются с одним центральным словарем в данной совокупности. Разнообразие структур обеспечивается возможностью устанавливать каждый раз соответствие либо в одну сторону, либо в обе. Отмечено, что наличие установленного соответствия от словаря А к словарю В не может служить основанием для автоматического вывода обратных соответствий от словаря В к словарю А; требуется независимое рассмотрение значений терминов в контексте каждого словаря. Отмечается также, что на практике чаще встречаются не эти три идеальные структуры, а их различные комбинации, и что нет необходимости во всех случаях стремиться к полному двухстороннему соответствию всех словарей. Указаны условия применимости разных структур, но обзор не обладает полнотой рассмотрения.

В главе 7 перечислены три типа соответствий, которые могут быть установлены для пары терминов (иерархическое отношение, эквивалентность и ассоциация), и указано, что предпочтительным является соответствие эквивалентности.

Глава 8 посвящена описанию различных видов соответствий типа эквивалентности. Вводятся понятия и обозначения точной и неточной эквивалентности, эквивалентности объединению понятий и эквивалентности пересечению понятий. Указано на необходимость учитывать контекст словаря при установлении соответствия эквивалентности даже при лексическом тождестве сопоставляемых терминов.

Иерархические соответствия рассмотрены в главе 9. Вводятся понятия и обозначения родовидовой, паритивной и инстанциальной иерархии. Указано на приоритетность установления родовидовых соответствий. Описание и обозначения точной, неточной и частичной иерархии отнесены к главе 11.

В главе 10 рассмотрено ассоциативное соответствие. Отмечается его близость к неточной эквивалентности и иерархии, что требует использования четких критериев для их различия.

Применение установленных соответствий для информационного поиска рассмотрено в главе 12 на семи примерах, показывающих преобразования поисковых образов документа (индексных терминов) и поисковых предписаний (поисковых терминов) при различных этапах соответствий – простая точная и неточная эквивалентность, эквивалентность пересечению и объединению, широкое и узкое иерархическое соответствие, ассоциация. Указана мера влияния типа соответствий на полноту и точность поиска. Сформулированы некоторые требования к организации работы по установлению соответствий в зависимости от целей использования. Ука-

зано на необходимость человеческого контроля в большом числе случаев при компьютерной реализации поиска с использованием соответствий.

Отдельная глава 13 посвящена обработке предкоординированных терминов, отражающих сложные понятия, которым соответствуют простые понятия в сопоставляемом словаре (тезаурусе). Указан порядок действий по установлению соответствий для перехода от сложных понятий к простым и наоборот. Приведены многочисленные примеры.

В главе 14 прописаны действия и приведены примеры сопоставления терминов в разных случаях применительно к ручной технологии и при автоматизации процессов. Глава 15 содержит рекомендации по оформлению и хранению данных разработанных сопоставительных таблиц. Указана необходимость тщательного ведения данных о соответствии при внесении изменений в сопоставляемые словари, приводятся сведения о необходимых корректировках. Визуализация данных о сопоставлениях рассмотрена в главе 16, где указано, что ни один из способов не удовлетворяет всем требованиям, которые могут возникнуть в частных случаях. При сопоставлении тезаурусов удобно указывать соответствия в дескрипторных статьях отдельными строками дополнительно к обычным тезаурусным элементам данных. При сопоставлении классификационных систем удобно составлять сопоставительные таблицы.

С главы ***17 Классификационные системы*** начинается рассмотрение отдельных типов словарей. В этой главе фактически обсуждается только один тип – библиографические классификации, т.е. системы, предназначенные для индексирования тематики информационных ресурсов. К этому типу относятся в частности УДК, ББК, Десятичная классификация Дьюи, Классификация Библиотеки Конгресса США, Классификация двоеточием Ранганатана, Библиографическая классификация Блисса.

Рассмотрение в этой и следующих главах идет по типовым рубрикам: общее описание, место в информационном поиске, словарный контроль, подвиды словарей данного типа, сопоставление смысловых единиц, рекомендации по сопоставлению с тезаурусами. Нормативное значение имеет только последняя рубрика, а предпоследняя важна постольку, поскольку она уточняет объект нормирования, и тут также даются некоторые рекомендации.

Классификационные системы, используемые в контексте ведения документооборота, рассмотрены в главе 18 по тем же типовым рубрикам. Главное своеобразие этих классификаций заключается в том, что их классы имеют не предметный характер, а обозначат виды деловой активности. В нормативной части главы даются рекомендации не столько по установлению соответствий сколько по применению соответствий в других информационных процессах.

В главе ***19 Таксономии*** идет речь о перечнях разделов электронных информационных ресурсов. В отечественной практике такие перечни обычно именуются как ***рубрикации***. Здесь обсуждаются такие вопросы как различие моноиерархических и полииерархических структур, языковая асимметрия значений терминов и другие, которые в равной степени приложимы ко всем классификационным системам. В нормативной части главы новым является указание на многообразие видов того,

что названо термином «типология», и необходимость вследствие этого тщательно подходить к анализу истинной природы рассматриваемого словаря. Здесь интерес представляет разбор семи примеров различных типов сопоставления.

Словари предметных рубрик, рассмотренные в главе 20, сочетают в себе свойства тезаурусов и классификационных систем. Как тезаурусы они представляют свои понятия в лексической форме, а как синтетические классификации они предусматривают образование предкоординированных индексов для сложных понятий. К этому типу относится словарь «Предметные рубрики Библиотеки Конгресса». Сопоставление простых рубрик таких словарей следует тем же правилам, что и сопоставление тезаурусов, описанным в главах 7–11. Здесь, в главе 20, сформулированы рекомендации по членению сложных рубрик на простые для сопоставления с дескрипторами тезауруса.

Глава 21 посвящена онтологиям – наиболее современному подходу к представлению предметной области информационных систем. Согласно классическому определению под это понятие в информатике подводятся все способы структуризации и поиска знаний, в частности и классификации и тезаурусы, но здесь рассматриваются именно «полноформатные» онтологии, состав которых разъясняется в ряде пунктов, описывающих следующие компоненты онтологий:

– классы (связаны родовидовыми отношениями подобно классам в классификациях и дескрипторам в тезаурусах);

– свойства (соответствуют выделению различных категорий дескрипторов и отношениям между членами одного класса и членами других классов);

– аксиомы (утверждения, определяющие основные качества классов, свойств и других сущностей в онтологии);

– индивиды (объекты рассмотрения, относительно которых онтология делает определенные утверждения);

– утверждения (особая группа аксиом, которые являются сообщениями о свойствах индивидов).

В качестве примера приведена простая онтология из области астрономии, состоящая из двух объектов и пяти свойств. Более состоятельным примером может служить онтология CIDOC CRM, получившая статус международного стандарта ИСО 21127:2014 [4].

Эти разделы стандарта восполняют отсутствие нормативного описания структуры онтологий подобного тому как стандарты ISO 25964-1, ГОСТ Р 7.0.91 и ГОСТ 7.25 описывают структуру информационно-поисковых тезаурусов. Нормативное содержание главы включает также рекомендации по взаимодействию онтологий с тезаурусами при четырех сценариях, и предусматривает как навигацию по ресурсам, доступным либо через онтологию, либо через тезаурус, так и разработку онтологии на основе имеющегося тезауруса.

Заголовок главы ***22 «Терминосистемы»*** выбивается из ряда наименований типов словарей, но по существу в ней идет речь о сопоставлении с терминологическими словарями, концепция которых определяется международным стандартом ISO 704:2009 [5]. В главе констатировано отличие тезаурусов в назначении и оформле-

нии, но большое сходство на уровне логической организации понятий. Рекомендовано использовать при сопоставлении те же методы, что и при сопоставлении двух тезаурусов, а результаты сопоставления использовать для расширения имеющегося тезауруса, который потом можно будет использовать при поиске в области, охватываемой терминосистемой.

В главе ***23 Авторитетные файлы имен*** рассматриваются нормативные списки наименований индивидуальных сущностей, которые подобно тезаурусам предназначены для индексирования документов. Сопоставление с другими словарями индексирования могут быть только частичными, поскольку авторитетные файлы включают имена только индивидуальных объектов, а тезаурусы и другие словари содержат главным образом имена классов, где индивиды (одноэлементные классы) составляют незначительное меньшинство. После подробного описания сущности и структуры авторитетных файлов имен в сопоставлении с тезаурусами даны краткие общие рекомендации по сопоставлению и приведен довольно подробный разбор семи примеров.

Последняя глава ***24 Словари синонимических рядов*** рассматривает словари синонимов, где для каждого входного термина даются списки выражений, которые в каких-либо контекстах могут иметь то же значение, что и заглавный термин (называемые синонимическими рядами или синсетами). После рассмотрения сущности синонимических рядов стандарт констатирует целесообразность использование их для расширения тезауруса системы при методах установления соответствий идентичных методам сопоставления тезаурусов.

Выводы

Проект стандарта описывает широкий спектр словарей, взаимодействие которых можно использовать для поиска информации в среде разнородных информационных ресурсов. Рассмотрены следующие типы словарей: информационно-поисковые тезаурусы, классификационные системы различного назначения, словари предметных рубрик, онтологии, терминологические словари, словари синонимов и авторитетные файлы имен. Определены смысловые соответствия, которые целесообразно устанавливать между значащими элементами словарей при их сопоставлении для взаимодействия. Установлены обозначения соответствий. Даны рекомендации по сопоставлению с тезаурусами каждого вида рассмотренных словарей.

Стандарт имеет характер методического пособия для разработчиков информационных систем и не столько нормирует методы взаимодействия лингвистических средств, сколько разъясняет целесообразность применения тех или иных приемов. Изложение стандарта следует традиции международной стандартизации, которая в отличие от российской традиции допускает в стандартах излишнее количество комментариев, не имеющих нормативного значения. Стандарт будет полезен именно в качестве такого пособия, хотя многие его рекомендации представляются фрагментарными и недостаточно проработанными. Стандарт будет полезен также как источник согласованной терминологии для описания словарей, основанных на различных концепциях представления предметной области информационного поиска.

Список литературы

1. *ISO 25964-2:2014* Information and documentation – Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies. Geneva: ISO, 2014.

2. *ISO 25964-2:2013* Information and documentation – Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. Geneva: ISO; 2011.

3. ГОСТ Р *7.0.91–2015* Система стандартов по информации, библиотечному и издательскому делу. Тезаурусы для информационного поиска. М.: Стандартинформ; 2015.

4. *ISO 21127:2014* Information and documentation – A reference ontology for the interchange of cultural heritage information. Geneva: ISO; 2014.

5. *ISO 704:2009* Terminology work – Principles and methods. Geneva: ISO; 2014.

Информация об авторах

Белоозеров Виктор Николаевич – кандидат филологических наук, доцент, ведущий научный сотрудник, Научно-методологическое отделение, Отдел развития классификационных систем, Всероссийский институт научной и технической информации РАН, г. Москва, Россия; ORCID: https://orcid.org/0000-0002-4200-1410, e-mail: nomoip@viniti.ru.

Антошкова Ольга Александровна – заместитель заведующего отделением, Научно-методологическое отделение, Всероссийский институт научной и технической информации РАН, г. Москва, Россия; e-mail: oant@viniti.ru.

Information about the authors

Viktor N. Beloozerov – Candidate of Science (Philology), Associate Professor, Leading Researcher, Division for Classification Systems Development, the Scientific and Methodological Department, All-Russian Institute for Scientific and Technical Information of the Russian Academy of Science (VINITI RAS), Moscow, Russia; ORCID: https://orcid.org/0000-0002-4200-1410, e-mail: nomoip@viniti.ru.

Olga A. Antoshkova – Deputy Head of the Scientific and Methodological Department, All-Russian Institute for Scientific and Technical Information of the Russian Academy of Science (VINITI RAS), Moscow, Russia; e-mail: oant@viniti.ru.