ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р 7.0.106— 2024 (ИСО 25964-2:2013)

Система стандартов по информации, библиотечному и издательскому делу

ВЗАИМОДЕЙСТВИЕ ТЕЗАУРУСОВ И ДРУГИХ СЛОВАРЕЙ

(ISO 25964-2:2013, Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies, MOD)

Издание официальное

Москва
Российский институт стандартизации
2024

gost.e-ecolog.ru

Предисловие

1 ПОДГОТОВЛЕН Федеральным государственным бюджетным учреждением «Государственная публичная научно-техническая библиотека России» (ГПНТБ России), на основе собственного перевода на русский язык англоязычной версии стандарта, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 191 «Научно-техническая информация, библиотечное и издательское дело, управление документами»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 29 февраля 2024 г. № 266-ст

4 Настоящий стандарт является модифицированным по отношению к международному стандарту ИСО 25964-2:2013 «Информация и документация. Тезаурусы и взаимодействие с другими словарями. Часть 2. Взаимодействие с другими словарями» (ISO 25964-2:2013 «Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies», MOD) путем:

- исключения информационного приложения, а также отдельных фраз, которые нецелесообразно применять в российской национальной стандартизации;
- включения дополнительных положений, фраз, ссылок, примечаний, сносок для учета особенностей российской национальной стандартизации, которые выделены в тексте курсивом;
- изменения отдельных фраз, которые выделены в тексте полужирным курсивом с подчеркиванием.

Наименование настоящего стандарта изменено относительно наименования указанного международного стандарта для приведения в соответствие с ГОСТ Р 1.5—2012 (пункт 3.5).

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном стандарте, приведены в дополнительном приложении ДА

5 ВВЕДЕН ВПЕРВЫЕ

*Правила применения настоящего стандарта установлены в статье 26 Федерального закона от 29 июня 2015 г. № 162-ФЗ «О стандартизации в Российской Федерации». Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомления и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (*gost.e-ecolog.ru

© ISO, 2013
© Оформление. ФГБУ «Институт стандартизации», 2024

Настоящий стандарт не может быть полностью или частично воспроизведен, тиражирован и распространен в качестве официального издания без разрешения Федерального агентства по техническому регулированию и метрологии

Содержание

1 Область применения 1
2 Нормативные ссылки 1
3 Термины и определения 2
4 Символы, сокращения и условные обозначения 11
5 Задачи и идентификация 12
5.1 Задачи совместимости и соответствия 12
5.2 Идентификация элементов соответствия 12
6 Структурные модели соответствия между словарями 13
6.1 Общие положения 13
6.2 Модель 1: единая структура 13
6.3 Модель 2: прямая связь 13
6.4 Модель 3: централизованная структура 14
6.5 Селективное соответствие 15
6.6 Выбор модели соответствия 16
7 Типы соответствий 17
7.1 Общие положения 17
8 Эквивалентные соответствия 17
8.1 Общие положения 17
8.2 Простая эквивалентность 17
8.3 Сложная эквивалентность 18
9 Иерархические соответствия 20
10 Ассоциативные соответствия 21
11 Точная, неточная и частичная эквивалентность 22
11.1 Общие положения 22
11.2 Точная эквивалентность 22
11.3 Неточная эквивалентность 22
11.4 Частичная эквивалентность 23
12 Использование соответствий в информационном поиске 24
12.1 Общие положения 24
12.2 Участие человека 24
12.3 Примеры реализации соответствий 24
12.4 Выводы и рекомендации 26
13 Управление предкоординацией 27
13.1 Общие положения 27
13.2 Сопоставление тезауруса и системы с контекстно-зависимыми рубриками 28
13.3 Соответствия более сложных классов 32
14 Методы установления соответствий 33
14.1 Общая процедура 33
14.2 Автоматизация прямого сопоставления 35
14.3 Соответствия на основе совместной встречаемости 35
14.4 Другие методы 35
15 Управление данными 36
15.1 Типы данных 36
15.2 Хранение данных 37
15.3 Сохранение данных о соответствиях 38
16 Визуализация сопоставленных словарей 39
16.1 Общие положения 39
16.2 Представление одной словарной статьи 40
16.3 Полное представление на базе одного из словарей 42
16.4 Сопоставительные таблицы 43
17 Классификационные системы 44
17.1 Ключевые характеристики и происхождение 44
17.2 Семантические компоненты и отношения в сопоставлении с компонентами тезауруса 46
17.3 Рекомендации по сопоставлению тезауруса и классификационной системы 49

18 Классификационные системы для управления документами 49
18.1 Ключевые характеристики и происхождение 49
18.2 Семантические компоненты и отношения в сопоставлении с компонентами тезауруса 51
18.3 Рекомендации по совместимости с тезаурусом 51
19 Таксономии 52
19.1 Ключевые характеристики и происхождение 52
19.2 Типы таксономий 53
19.3 Семантические компоненты и отношения 54
19.4 Сопоставление тезауруса и таксономии 55
20 Словари предметных рубрик 59
20.1 Ключевые характеристики и происхождение 59
20.2 Семантические компоненты и отношения 60
20.3 Сопоставление предметных рубрик с тезаурусными понятиями 61
21 Онтологии 64
21.1 Ключевые характеристики и происхождение 64
21.2 Семантические компоненты и отношения 65
21.3 Структурное сравнение тезаурусов и онтологий 68
21.4 Совместимость с тезаурусами 68
22 Терминосистемы 70
22.1 Ключевые характеристики и происхождение 70
22.2 Семантические компоненты и отношения в сравнении с аналогичными компонентами и отношениями тезауруса 71
22.3 Совместимость с тезаурусами 72
23 Нормативные файлы имен 72
23.1 Ключевые характеристики и происхождение 72
23.2 Семантические компоненты и отношения 74
23.3 Сопоставление тезауруса и нормативного списка имен 75
24 Словари синонимических рядов 78
24.1 Ключевые характеристики и происхождение 78
24.2 Семантические компоненты и отношения 78
24.3 Взаимодействие с тезаурусами 79
Приложение ДА (справочное) Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном стандарте 80
Библиография 81

Введение

Способность идентифицировать и находить необходимую информацию среди обширных коллекций и других ресурсов является одной из основных и насущных проблем сегодня, поэтому возникает необходимость в семантической совместимости ресурсов. Чтобы удовлетворить эту потребность, активно разрабатывают различные веб-сервисы и другие инструменты, включая следующие:

- реестры словарей и схем метаданных,
- репозитории словарей и схем метаданных,
- сопоставительные таблицы словарей.

Настоящий стандарт содержит указания по разработке сопоставительных таблиц, а реестры и репозитории находятся вне сферы рассмотрения. Сопоставительные таблицы указывают соответствия элементов различных схем и словарей. Межсловарные соответствия являются основным объектом этого стандарта.

Основа для рассмотрения функциональной совместимости словарей заложена в ГОСТ Р 7.0.91, который описывает ключевые характеристики тезаурусов и дает рекомендации по их разработке и ведению. К сожалению, для других типов словарей аналогичных стандартов не существует. По этой причине в настоящем стандарте содержится некоторое элементарное описание других словарей, таких как классификационные системы, словари предметных рубрик и т. п.

Разделы 5—12 и 14—16 настоящего стандарта касаются принципов и практических возможностей сопоставления, которые применимы к большинству типов словарей и особенно к тезаурусам. В разделе 13 приведены дополнительные рекомендации по обработке предкоординированных классов, которые встречаются в системах классификации и других словарях, использующих классификационный подход.

Разделы 17—24 посвящены различным типам словарей. Первоочередное внимание уделяется словарям, которые обычно используются для классификации или индексирования информационных ресурсов, а именно системам классификации (в том числе используемым для управления документами), таксономиям, словарям предметных рубрик и нормативным файлам имен. Также включены терминологические словари, онтологии и словари синонимов, несмотря на отличия в их назначении. В каждом разделе приведено краткое информативное описание ключевых характеристик словаря в сравнении его семантических компонент с тезаурусом, чтобы обеспечить понимание рекомендаций по сопоставлению.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Система стандартов по информации, библиотечному и издательскому делу

ВЗАИМОДЕЙСТВИЕ ТЕЗАУРУСОВ И ДРУГИХ СЛОВАРЕЙ

System of standards on information, librarianship and publishing. Interoperability of thesauri and other vocabularies

Дата введения — 2024—05—01

1 Область применения

В настоящем стандарте описаны методы сопоставления структуры и элементов тезаурусов и других типов словарей, которые обычно используются для поиска информации. В нем описываются, сравниваются и противопоставляются элементы и особенности этих словарей, которые влияют на совместимость.

Настоящий стандарт дает рекомендации по созданию и поддержанию соответствий между несколькими тезаурусами или между тезаурусами и другими типами словарей.

Стандарт предназначен для любых организаций, использующих в своей информационной практике соответствия между тезаурусами или между тезаурусами и другими словарями.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ГОСТ 7.24 Система стандартов по информации, библиотечному и издательскому делу. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению

ГОСТ 7.59 Система стандартов по информации, библиотечному и издательскому делу. Индексирование документов. Общие требования к систематизации и предметизации

ГОСТ 7.75 Система стандартов по информации, библиотечному и издательскому делу. Коды наименований языков

ГОСТ Р 7.0.66 (ИСО 5963:1985) Система стандартов по информации, библиотечному и издательскому делу. Индексирование документов. Общие требования к координатному индексированию

ГОСТ Р 7.0.91—2015 (ИСО 25964-1:2011) Система стандартов по информации, библиотечному и издательскому делу. Тезаурусы для информационного поиска

ГОСТ Р ИСО 704—2010 Терминологическая работа. Принципы и методы

ГОСТ Р ИСО 15489-1 Система стандартов по информации, библиотечному и издательскому делу. Информация и документация. Управление документами. Часть 1. Понятия и принципы

Примечание — При пользовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указателю «Национальные стандарты», который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя «Национальные стандарты» за текущий год. Если заменен ссылочный стандарт, на который дана недатированная ссылка, то рекомендуется использовать действующую версию этого стандарта с учетом всех внесенных в данную версию изменений. Если заменен ссылочный стандарт, на который дана датированная ссылка, то рекомендуется использовать версию этого стандарта с указанным выше годом утверждения (принятия). Если после утверждения настоящего стандарта в ссылочный стандарт, на который дана датированная ссылка, внесено изменение, затрагивающее положение, на которое дана ссылка, то это положение рекомендуется применять без учета данного изменения. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, рекомендуется применять в части, не затрагивающей эту ссылку.

Издание официальное

3 Термины и определения

В настоящем стандарте применены следующие термины с соответствующими определениями:

3.1 **классификационный ряд** (array): Группа соподчиненных понятий (непосредственно подчиненных одному обобщающему понятию).

Пример — Соподчиненные понятия «верхняя одежда» и «нижняя одежда» формируют классификационный ряд в рамках понятия «одежда»:

Одежда
Верхняя одежда
Пальто
Нижняя одежда

3.2 **ассоциативное отношение** (associative relationship): Отношение между двумя понятиями, не связанными иерархически, но имеющими сильную смысловую связь.

3.3 **вышестоящий термин** (broader term): Предпочтительный термин (дескриптор), обозначающий более широкое понятие, чем обозначаемое данным дескриптором.

Примечание — Тематическая область более узкого понятия полностью входит в тематическую область более широкого понятия. Отношения между этими двумя понятиями обычно обозначаются ***меткой ВТ в английском языке и меткой «в» (выше) в русском языке. Более подробные объяснения см. в ГОСТ Р 7.0.91—2015, пункт 10.2.1.***

3.4 **наименование класса** (caption): Текстовая метка класса, представленного в классификационной системе классификационным кодом.

Примечание — Наименование класса имеет значение, учитывающее иерархический контекст. Оно не обязано быть настолько же полным или таким же самодостаточным, как лексическое примечание или даже предпочтительный термин в тезаурусе. Наименования классов иногда называют классификационными рубриками или именами классов.

3.5 **категория** (category): Понятие или группа сходных или родственных понятий, используемая как раздел или подраздел в таксономии.

Примечания

1 В классификационных системах подобные группы обычно называют классами.

2 Это определение категории не следует путать с «фундаментальными категориями», которые используются в ***ГОСТ Р 7.0.91—2015 (например, в разделе 12)*** в контексте фасетной классификации.

3.6 **категориальная метка** (category label): Текстовая метка, представляющая категорию в таксономии.

Примечание — Категориальная метка в таксономии сопоставима с наименованием класса в классификационной системе и, хотя категорию часто называют «узел», категориальную метку не следует путать с меткой узла.

3.7 **цепной указатель** (chain index): Указатель к схеме, представляющий сложные понятия с помощью строки предкоординированных терминов, например в классификационной системе, в которой записи формируются путем последовательного усечения слева строки терминов, представляющих сложное понятие.

Примечание — См. пример в 17.2.4, рисунок 12.

3.8 **основание деления** (characteristic of division): Признак, по которому понятие может быть разделено на ряд более узких понятий, каждое из которых имеет значение этого признака, отличное от других.

См. термин «узловая метка».

Пример — Возрастная группа является основанием деления понятия «люди»:
Люди
(по возрасту)
Дети
Молодежь
Взрослые

3.9 **фасетная формула, порядок следования** (citation order): Порядок, в котором предпочтительные термины или классификационные коды объединяются в системе предкоординатного индексирования или в классификационной системе для формирования строк, представляющих сложные понятия.

3.10 **класс** (class): Понятие или группа сходных или родственных понятий, используемая как раздел или подраздел в классификационной системе.

Примечание — Классы — это основные единицы, из которых построена классификационная система. В таксономиях, которые являются разновидностью классификационных систем, они обычно известны как категории. Иногда их называют «узлы», но их не следует путать с метками узлов. Термин «класс» в контексте онтологий имеет другое значение. См. 21.2.2.

3.11 **классификация** (classification), **классифицирование** (classifying): Деятельность, подразумевающая объединение сходных или родственных объектов вместе, разделение несходных или неродственных объектов и представление результирующих классов в логической и удобной последовательности.

3.12 **классификационная система** (classification scheme): Система понятий и предкоординированных комбинаций понятий, организованная классификацией.

Примечание — Классификационные системы часто включают указатели понятий.

3.13 **коллекция** (collection): Набор информационных ресурсов, к которым может быть получен доступ из структурированного словаря, независимо от того, собраны ли элементы набора в одном месте или распределены в сети.

3.14 **сложная эквивалентность** (compound equivalence): Отношение между терминами или соответствие между понятиями, в рамках которого один термин или понятие одного контекста представлено двумя или более терминами или понятиями в другом контексте.

3.15 **сложная рубрика** (compound heading), **предкоординированная рубрика** (pre-coordinated heading): Предметная рубрика, образованная предкоординацией более чем одного термина в одну строку.

Пример — Отдельные термины «Психология личности», «Религиозный аспект» и «История» могут быть объединены в сложную рубрику «Психология личности — Религиозный аспект — История».

3.16 **составной термин** (compound term): Термин, который в соответствии с морфологией может быть разложен на самостоятельные отдельные компоненты.

Примеры
1 В английском языке:
«coppermines» можно разложить на «copper» и «mines»
«lawnmowers» можно разложить на «lawns» и «mowers»
2 Во французском языке:
«mine de cuivre» можно разложить на «mine» и «cuivre»
«biodiversité» можно разложить на «biologie» и «diversité»
3 В русском языке:
«медный рудник» можно разложить на «медь» и «рудник»
«биоразнообразие» можно разложить на «биология» и «разнообразие»

Примечание — Составной термин может состоять как из нескольких слов, так и из одного слова.

3.17 **понятие** (concept): Объект мышления.

Примечание — Понятия могут выражаться разными способами. Они существуют в сознании в виде абстрактных сущностей, которые независимы от терминов, используемых для их выражения. Они варьируют от очень простых понятий, например, «ребенок», до очень сложных, например, «законодательство о защите детей».

3.18 **группа понятий** (concept group): Совокупность понятий, выбранных по некоторому заданному критерию, например, понятия, имеющие отношение к конкретной предметной области.

3.19 **контролируемый словарь** (controlled vocabulary): Нормативный список терминов, рубрик или кодов, каждый из которых представляет понятие.

Примечание — Контролируемые словари предназначены для приложений, в которых полезно идентифицировать каждое понятие с помощью одной и той же рубрики, например, при классификации документов, их индексировании и/или поиске. Тезаурусы, словари предметных рубрик и нормативные файлы имен являются примерами контролируемых словарей.

См. термин «терминосистема».

3.20 **межъязыковая эквивалентность** (cross-language equivalence): Отношение эквивалентности между терминами, представляющими одно и то же понятие в различных языках.

3.21 **сопоставительная таблица** (crosswalk): Таблица соответствий между понятиями в двух или более структурированных словарях.

Примечание — Сопоставительные таблицы обеспечивают способность поисковых систем выполнять поиск по разнородным базам данных.

3.22 **модель данных** (data model): Абстрактная модель, описывающая то, как данные представляются и используются.

Примечание — *Модель данных в ГОСТ Р 7.0.91 обеспечивает общее определение структуры и семантики тезауруса. Она может быть использована в качестве основы для определения модели базы данных и обменного формата тезаурусов.*

3.23 **дифференцированное сопоставление** (differentiated mapping): Методика сопоставления, предусматривающая установление различий между соответствиями разных типов и характеристик.

Примечание — Типы соответствий, которые можно различить, включают эквивалентность, ассоциативность и иерархичность. Эквивалентность может быть далее подразделена на простую и сложную. Для обеспечения дополнительных различий может быть указана степень эквивалентности.

См. термин «недифференцированное сопоставление».

3.24 **документ** (document): Любой ресурс, который может быть классифицирован или индексирован для того, чтобы стал возможным поиск содержащихся в нем данных или информации.

Примечание — Это определение распространяется не только на материалы, написанные и напечатанные на бумажном носителе или представленные в виде микрофильма (обычные книги, журналы, диаграммы, карты), но и на непечатные средства передачи информации (машиночитаемые и оцифрованные записи, ресурсы Интернета и интранета, фильмы, звукозаписи, люди и организации как носители научных знаний, здания, местности, монументы, трехмерные объекты), а также коллекции таких единиц и их составные части.

3.25 **входной термин** (entry term), **вводящий термин** (lead-in term): Термин, представленный в контролируемом словаре, но используемый не непосредственно в качестве метаданных, а для того, чтобы привести пользователя к другому термину, имеющему статус категориальной метки, предметной рубрики или дескриптора.

Примечание — Входной термин в составе тезауруса часто называют непредпочтительным термином и аскриптором.

3.26 **перечислительная классификационная система** (enumerative classification scheme): Классификационная система, в которой все доступные классы перечислены в таблицах.

См. термин «синтетическая классификационная система».

3.27 **эквивалентное соответствие** (equivalence mapping): Соответствие, фиксирующее некое понятие в целевом словаре, которое признается идентичным по объему понятию исходного словаря.

См. термин «отношение эквивалентности».

3.28 **отношение эквивалентности** (equivalence relationship): Отношение между двумя терминами в тезаурусе, показывающее, что оба эти термина обозначают одно и то же понятие.

Примечание — В обычном словоупотреблении термины, являющиеся квазисинонимами, могут представлять слегка различающиеся понятия. Однако включение в тезаурус устанавливаемого между ними отношения эквивалентности определяет, что оба эти термина рассматриваются как представители одного и того же понятия. Когда в одноязычный или многоязычный тезаурус включены два или более термина одного и того же языка, то один из них выбирается в качестве предпочтительного термина (дескриптора), а другой (другие) — в качестве непредпочтительного термина (аскриптора). Когда два или более таких терминов являются представителями разных языков в многоязычном тезаурусе, каждый из них может выступать как дескриптор в своем собственном языке, и такие отношения принято называть межъязыковой эквивалентностью.

3.29 **обменный формат** (exchange format): Машиночитаемый формат для представления информации, предназначенный для облегчения обмена информацией между различными приложениями.

Примечание — Обменный формат для тезауруса часто использует язык разметки, например, на основе стандарта XML (Extensible Markup Language), и основывается на модели данных тезауруса. Если модель данных

представляет собой общее описание структуры и семантики тезауруса, то обменный формат выражает это на формальном языке с целью обмена тезаурусами.

3.30 **фасет** (facet): Группа однородных понятий одной и той же природной категории.

Примеры

1 Животные, мыши, нарциссы и бактерии могут рассматриваться как члены фасета живых организмов.

2 Копание, писание и приготовление еды могут рассматриваться как члены фасета действий.

3 Париж, Великобритания и Альпы могут рассматриваться как члены фасета территорий.

Примечание — Примерами таких категорий высокого уровня, которые могут быть использованы для группировки понятий в фасеты, являются следующие категории: предмет, материал, действующий фактор, действие, место и время.

См. термин «узловая метка».

3.31 **фасетная классификационная система** (faceted classification scheme): Классификационная система, в которой понятия анализируются на основании их фасетов.

Примечание — Классификационные таблицы составляются для каждого фасета, а затем термины или классификационные коды из них объединяются в соответствии с предписанными правилами для выражения сложных понятий. Некоторые сложные понятия часто перечисляются в классификационных таблицах; другие могут быть синтезированы пользователем.

3.32 **иерархическое отношение** (hierarchical relationship): Отношение между двумя понятиями, при котором объем одного из них полностью находится внутри объема другого.

Примечание — Существуют несколько разных типов иерархических отношений. *Более подробно см. ГОСТ Р 7.0.91—2015*, подраздел 10.2.

См. термины «вышестоящий дескриптор», «нижестоящий дескриптор».

3.33 **омограф** (homograph): Одно из двух или более слов, которые пишутся одинаково, но имеют разное значение.

Примеры

1 В английском языке слово «bank» может означать и финансовую организацию, и берег реки.

2 Во французском языке слово «avocat» может означать и юриста, и фрукт.

3 В русском языке слово «ключ» может означать и родник, и инструмент.

Примечание — *Омографы отличаются от омофонов, т. е. таких пар слов, как «weights» и «waits» в английском, или «mer» и «mère» во французском, или «кампания» и «компания» в русском языке, которые пишутся по-разному и имеют разное значение, а читаются одинаково.*

3.34 **идентификатор** (identifier): Набор знаков, обычно алфавитно-цифровых, обозначающий понятие, термин или какую-то другую сущность и используемый для достижения однозначной идентификации внутри определенного контекста или ресурса, особенно в компьютерных системах или сетях.

Примечание — Иногда в качестве идентификатора используют классификационный код.

3.35 **индексный термин** (index term): Термин, приписанный документу в процессе индексирования.

Примечание — Иногда индексные термины называют терминами индексирования, ключевыми словами или метками. Но два последних термина являются многозначными. В качестве индексных терминов часто используются дескрипторы тезаурусов.

3.36 **индексирование** (indexing): Интеллектуальный анализ предметного содержания документа для идентификации представленных в нем понятий и предоставление соответствующих индексных терминов для обеспечения поиска информации.

Примечание — *Для обозначения этого действия используется термин «предметное индексирование», но, поскольку в стандартах на индексирование (ГОСТ 7.59, ГОСТ Р 7.0.66, ГОСТ Р 7.0.91) такие «непредметные» элементы, как имена авторов и даты не рассматриваются, достаточно использовать краткий термин «индексирование». Индексирование может осуществляться как экспертами, так и в автоматическом режиме.*

3.37 **информационный поиск** (information retrieval): Все методы и процессы, используемые для того, чтобы выбрать из коллекции или сети информационных ресурсов документы, отвечающие информационным потребностям.

Примечание — Это определение включает отбор и включение документов в выборку, а также их просмотр и другие формы поиска информации.

3.38 **совместимость** (interoperability): Способность двух или более систем или модулей обмениваться информацией и использовать информацию, полученную в результате такого обмена.

Примечание — Словари могут поддерживать совместимость путем указания соответствий с другими словарями, путем представления информации в стандартных форматах и путем использования систем, поддерживающих общие компьютерные протоколы.

3.39 **сопоставить** (map, *verb*): Установить отношения между понятиями одного словаря и понятиями другого словаря.

3.40 **сопоставление (процесс)** (mapping, *gerund*): Процесс установления отношений между понятиями одного словаря и понятиями другого словаря.

3.41 **соответствие** (mapping, *noun*): Результат процесса сопоставления, т. е. отношение между понятием одного словаря и одним или более понятием другого словаря.

Примечания

1 Как правило, соответствие имеет направление, которое описано в разделе 6.

2 В исключительных случаях соответствие может включать комбинацию из двух или более целевых словарей, где один или несколько из них действуют как классификатор для другого (см. 8.3.4).

3.42 **кластер соответствий** (mapping cluster): Согласованный набор соответствий между понятиями трех или более словарей.

Примечание — Примеры кластеров соответствий приведены в 6.3 и 6.4. Кластер соответствий поддерживается и публикуется для применения в определенном приложении. Например, можно поддерживать кластер соответствий между четырьмя различными тезаурусами, чтобы пользователь любого из них мог легко выполнять поиск в коллекциях документов, проиндексированных по любому из них.

3.43 **разметка** (markup): Примечания или какие-либо виды кодирования, включенные в текст согласно правилам языка разметки.

3.44 **язык разметки** (markup language): Набор правил кодирования, которые применяются для составления инструкций по интерпретации текста путем использования примечаний, включенных непосредственно в сам текст.

Примечание — Интерпретация касается таких вопросов, как содержание, структура и представление текста. В качестве примеров широко используемых языков разметки можно привести HTML (Hypertext Markup Language), который в основном ориентирован на представление, и XML (Extensible Markup Language), который определяет структуру текста.

3.45 **метаданные** (metadata): Данные, которые идентифицируют атрибуты документа, используемые для поддержки функций размещения, обнаружения, документирования, оценки и/или выбора.

Примечание — В качестве значений метаданных применяют дескрипторы или классификационные коды, выбранные в процессе индексирования.

3.46 **микротезаурус** (microthesaurus): Выделенное подмножество тезауруса, способное функционировать как полный тезаурус.

3.47 **моноиерархическая структура** (monohierarchical structure): Иерархическая организация понятий в тезаурусе или классификационной системе, в которой каждое понятие может иметь непосредственно над собой только одно вышестоящее понятие.

См. термин «полииерархическая структура».

Пример — В моноиерархической структуре понятие «фортепиано» не может одновременно причисляться и к клавишным инструментам, и к струнным инструментам. Чтобы определить его место в структуре, следует выбрать одно из этих понятий.

3.48 **многоязычный тезаурус** (multilingual thesaurus): Тезаурус, в котором термины и структура отношений представлены на двух или более естественных языках.

3.49 **многословный термин** (multi-word term): Термин, состоящий более чем из одного слова.

См. термин «составной термин».

Пример — cost benefit analysis, анализ выгод и затрат.

3.50 **нормативный список имен** (name authority list): Контролируемый словарь, предназначенный для последовательного именования индивидуальных объектов.

Примечание — Эти объекты являются индивидуальностями, такими как Бенджамин Дизраэли, Килиманджаро или гобелен из Байе, а не являются классами, как политики, горы или вышивки. Нормативный список имен также известен как авторитетный файл имен. В настоящем стандарте нормативный список имен иногда называется просто нормативным списком.

3.51 **нижестоящий термин** (narrower term): Предпочтительный термин, представляющий более узкое понятие по сравнению с рассматриваемым термином.

Примечание — Область значений нижестоящего термина полностью располагается внутри области значений вышестоящего термина. Отношение между ними обычно обозначается ***меткой NT в английском языке и меткой «н» (ниже) в русском языке. Более подробно см. ГОСТ Р 7.0.91—2015, пункт 10.2.1.***

3.52 **узловая метка** (node label): Обозначение, проставляемое в иерархическом или классификационном указателе для того, чтобы показать, как упорядочены термины.

Примечание — Узловая метка не является ни дескриптором, ни аскриптором. Она содержит один из двух видов информации:

a) имя фасета, к которому принадлежат перечисленные ниже термины;

b) атрибут или основание деления, с помощью которого отсортирован или сгруппирован классификационный ряд близкородственных терминов. См. примеры в ***ГОСТ Р 7.0.91—2015, раздел 11.***

3.53 **непредпочтительный термин** (non-preferred term), **аскриптор** (non-descriptor): Термин, который не приписывают документу, а используют лишь для входа в тезаурус или указатель.

См. термин «входной термин».

Пример —

hounds	***пес***
USE dogs	***см. собака***

Примечание — За аскриптором (непредпочтительным термином) следует ссылка на соответствующий дескриптор (предпочтительный термин). В этом примере «hounds» и «пес» — аскрипторы, а «dogs» и «собака» — те дескрипторы, которые следует употреблять вместо них.

3.54 **классификационный код** (notation), **код класса** (class code), **индекс класса** (class number), **метка класса** (classmark): Набор знаков, представляющий понятие или класс в структурированном словаре, особенно в классификационной системе.

Пример —

Классификационный код	*Словарь-источник*	*Понятие*
07.04.4	***Тезаурус Международной организации труда (ILO)***	***Политика и развитие рыболовства***
622.342 2	***Десятичная классификация Дьюи***	***Добыча золота***
373.3.016:51	***Универсальная десятичная классификация***	***Курс математики в начальной школе***
SBS XEJ B	***Библиографическая классификация Блисса***	***Закон об охране исчезающих видов***
H40-H42	***Международная статистическая классификация болезней и проблем, связанных со здоровьем***	***Глаукома***

Примечание — Классификационный код иногда используют для того, чтобы отсортировать и/или разместить понятия в предопределенном систематическом порядке и показать, каким образом структурированы и сгруппированы компоненты сложных понятий. Классификационный код может осуществлять связь между алфавитным и систематическим указателями тезауруса. В контексте классификационных систем «понятия» часто называют «темами», особенно если они, подобно приведенным выше примерам, являются сложными.

3.55 **соответствие «один-ко-многим»** (one-to-many mapping): Соответствие, в котором одно понятие в одном словаре сопоставлено комбинации двух или более понятий в другом словаре.

Примечание — Эта ситуация отличается от той, в которой понятие имеет два или более независимых соответствия с понятиями в другом словаре.

3.56 **соответствие «один-к-одному»** (one-to-one mapping): Соответствие, в котором одно понятие в одном словаре сопоставлено одному понятию в другом словаре.

Примечания

1 Термины или классификационные коды, используемые для обозначения соответствующих понятий в двух словарях, могут быть или не быть идентичными.

2 У одного понятия может быть два или более однозначных соответствия, если они не зависят друг от друга.

3.57 **онтология** (ontology): Детальная формализация некоторой области знаний с помощью определенной концептуальной системы.

Примечание — Это определение дано Студером и др. [31] как развитие более раннего определения Грубера [18] и приведено здесь, поскольку широко распространено в сообществе разработчиков онтологий. Типичная онтология включает дефиниции понятий и отношений между ними, установленные формализованным образом в целях автоматизации логического вывода. Это определение исключает тезаурусы, классификационные системы и другие структурированные словари, описанные в настоящем стандарте, несмотря на то, что их иногда называют «облегченными онтологиями».

3.58 **полииерархическая структура** (polyhierarchical structure): Иерархическая организация понятий в тезаурусе или классификационной системе, в которой каждое понятие может иметь более одного вышестоящего понятия.

См. термин «моноиерархическая структура».

Пример — В полииерархической структуре понятие «орга́ны» (музыкальные инструменты) можно причислять и к клавишным инструментам, и к духовым.

Примечание — В полииерархической структуре одно и то же понятие может появляться более чем в одном месте иерархической структуры тезауруса. Его атрибуты и связи, и особенно нижестоящие и ассоциативные термины остаются неизменными вне зависимости от того, где термин встретился.

3.59 **посткоординация** (post-coordination): Комбинирование дескрипторов (предпочтительных терминов) контролируемого словаря в ходе поиска.

См. термин «предкоординация».

Пример — Посткоординированное поисковое предписание «микроволны и излучения» можно использовать для поиска документов о микроволновом излучении, если они были проиндексированы с помощью отдельных терминов «микроволны» и «излучения», а не с помощью составного термина.

3.60 **предкоординация** (pre-coordination): Комбинирование понятий, классов или терминов контролируемого словаря во время создания этого словаря или во время его использования для индексирования или классифицирования.

См. термин «посткоординация».

Примеры

1 Класс «Общая теория», когда он находится в составе более широкого класса «Музыка», соотносится только с предкоординированной темой «Теория музыки», а не с теорией вообще.

2 Предкоординированная цепочка «картон — переработка» может находиться в словаре предметных рубрик или (если она не включена в словарь) может быть синтезирована индексатором, когда окажется необходимой для индексирования конкретного документа.

3.61 **точность поиска** (precision): Показатель эффективности поиска, равный R/T, где R — количество полученных релевантных документов, а T — общее количество документов, выданных из одной коллекции.

3.62 **предпочтительный термин** (preferred term), **дескриптор** (descriptor): Термин, используемый для представления понятия при индексировании.

См. термин «аскриптор (непредпочтительный термин)».

Примечание — Дескриптор — это, как правило, существительное или именное словосочетание.

3.63 **протокол** (protocol): Соглашение, которое определяет синтаксис, семантику и синхронизацию процесса коммуникации между двумя компьютерами для решения определенных задач.

3.64 **квазисиноним** (quasi-synonym), **неполный синоним** (near-synonym): Один из двух или более терминов, значения которых в рамках обычного использования, как правило, рассматриваются как различные, но в данном контролируемом словаре они могут рассматриваться в качестве меток для одного и того же понятия.

Пример —

diseases, disorders	***болезни, недомогания***
earthquakes, earthtremors	***землетрясения, сейсмическая активность***

3.65 **полнота поиска** (recall): Показатель эффективности поиска, равный *R/N*, где *R* — количество полученных релевантных документов, а *N* — общее количество релевантных документов в коллекции.

3.66 **ассоциативный термин** (related term): Дескриптор (предпочтительный термин), обозначающий такое понятие, которое имеет ассоциативное отношение к рассматриваемому термину.

Примечание — Отношения между ассоциативными терминами обычно обозначаются ***меткой RT в английском языке и меткой «а» (ассоциация) в русском языке. Более подробно см. в ГОСТ Р 7.0.91—2015, подраздел 10.3.***

3.67 **классификационная таблица** (schedule): Совокупность терминов, классификационных кодов, рубрик, перекрестных ссылок и лексических примечаний, которая служит для представления содержания и структуры структурированного словаря.

3.68 **лексическое примечание** (scope note): Примечание, которое определяет или уточняет семантические границы понятия в рамках его использования в структурированном словаре.

Примечание — Термин, используемый для обозначения понятия, при обычном использовании может иметь несколько значений. Лексическое примечание применяют для закрепления за ним только одного из таких значений, и, где это необходимо, оно также указывает другие понятия, которые включены или исключены из объема уточняемого понятия.

3.69 **поисковый термин** (search term): Термин, образующий поисковый запрос или его часть.

Примечание — В контексте настоящего стандарта поисковые термины обычно выбирают из контролируемого словаря.

3.70 **поисковый тезаурус** (search thesaurus): Словарь, предназначенный для облегчения поиска, даже если он не использовался для индексирования совокупности документов, по которой выполняется поиск.

Примечание — Поисковые тезаурусы предназначены для облегчения выбора терминов и/или расширения поисковых предписаний путем включения терминов, обозначающих более широкие, более узкие или связанные понятия, а также синонимов. *В качестве поискового тезауруса можно использовать тезаурус, соответствующий ГОСТ Р 7.0.91.*

3.71 **исходный язык** (source language): Язык, служащий отправной точкой в процессе перевода или поиска эквивалентов для терминов.

3.72 **исходный словарь** (source vocabulary): Словарь, для терминов которого отыскивают соответствующие термины или понятия в другом словаре.

3.73 **специфичность словаря** (specificity): Способность структурированного словаря выразить предмет поиска углубленно и подробно.

Примечание — Более подробное рассмотрение специфичности словаря см. в ***ГОСТ Р 7.0.91—2015, подраздел 8.4 и др.***

3.74 **структурированный словарь** (structured vocabulary): Организованный набор терминов, рубрик или кодов, представляющих понятия и их взаимосвязи, который может быть использован для обеспечения информационного поиска.

Примечание — Структурированный словарь также может быть использован в других целях. В контексте поиска информации словарь нуждается в сопутствующих правилах, описывающих, как следует применять термины. В настоящем стандарте рассматриваются различные типы структурированных словарей, в том числе классификационные системы, словари предметных рубрик и др.

3.75 **предмет, тема** (subject): Понятие или сочетание понятий, рассматриваемое в документе или встречающееся в дискурсе.

3.76 **предметная рубрика** (subject heading): Термин или предкоординированная строка терминов, взятая из словаря предметных рубрик.

3.77 **словарь предметных рубрик** (subject heading scheme), **список предметных рубрик** (subject heading list), **язык предметных рубрик** (subject heading language), СПР (SHL): Структурированный словарь, состоящий из терминов, предназначенных для предметного индексирования, и правил для объединения их при необходимости в предкоординированные цепочки.

3.78 **синоним** (synonym): Каждый из двух (или более) терминов, обозначающих одно и то же понятие.

Примеры
1 В английском языке:
guarantees, warranties;
heart attack, myocardial infarction;
HIV, human immunodeficiency virus.
2 Во французском языке:
schiste, phyllade;
VIH, virus de l'immunodéficience humaine;
crise cardiaque, infarctus du myocarde.
3 В русском языке:
лингвистика, языковедение, языкознание;
ВИЧ, вирус иммунодефицита человека;
сердечный приступ, инфаркт миокарда.

Примечание — Сокращенная и полная формы термина могут рассматриваться как синонимы.

3.79 **синонимический ряд;** *кольцо синонимов, синсет* (synonym ring): Набор синонимичных или почти синонимичных терминов, любой из которых может использоваться для ссылки на одно и то же понятие.

Пример —
stream, river, brook, beck, burn;
водоток, река, речка, ручей, ручеек.

3.80 **синтетическая классификационная система** (synthetic classification scheme): Классификационная система, в которой пользователи могут синтезировать классификационные коды для сложных классов из списков более простых классов.

См. термин «перечислительная классификационная система».

3.81 **целевой язык** (target language): Язык, на котором выполнен перевод или найден эквивалент термина исходного языка.

3.82 **целевой словарь** (target vocabulary): Словарь, в котором ищется термин или понятие, соответствующее существующему термину или понятию в исходном словаре.

3.83 **таксономия** (taxonomy): Система категорий и подкатегорий, которые могут быть использованы для сортировки и иного упорядочения знаний или информации.

Примечание — Таксономии варьируют от очень простых до очень сложных. В простейших таксономиях категории не обязательно делятся на подкатегории, в то время как в сложных можно найти несколько иерархических уровней. Также могут присутствовать другие функции, такие как все функции тезауруса, описанные в ***ГОСТ Р 7.0.91***, и/или функции, обычно встречающиеся в классификационных системах. За пределами этого стандарта термин часто используется свободно для обозначения структурированного словаря любого типа.

3.84 **термин** (term): Слово или словосочетание, используемое для обозначения понятия.

Пример —

schools	***школы***
school uniform	***школьная форма***
costs of schooling	***плата за обучение***
teaching	***преподавание***

Примечание — Термины тезауруса могут быть как предпочтительными терминами (дескрипторами), так и непредпочтительными терминами (аскрипторами).

3.85 **терминосистема** (terminology): Набор обозначений, принадлежащих одному специальному языку.

Примечание — Термин «специальный язык (special language)» определяется в [3] как «язык, применяемый в определенной предметной области и характеризующийся использованием специальных языковых средств обозначения понятий»; а «обозначение (designation)» определяется как «представление понятия знаком, который на него указывает (denotes)».

3.86 **структурированный тезаурус, тезаурус** (thesaurus): Контролируемый структурированный словарь, в котором понятия представлены терминами, организованными таким образом, что отношения между понятиями четко выражены, и предпочтительные термины снабжены указателями перехода от синонимов и квазисинонимов.

Примечание — Тезаурус решает задачу обеспечения того, чтобы индексатор и пользователь выбирали для представления определенной темы (предмета) один и тот же дескриптор или комбинацию дескрипторов. По этой причине тезаурус оптимизирован как средство навигации и терминологического определения предметной области.

3.87 **наивысший термин** (top term): Дескриптор (предпочтительный термин), представляющий понятие, для которого в тезаурусе не существует более широкого понятия.

Примечание — Его иногда обозначают аббревиатурой TT (*в английском языке*).

3.88 **недифференцированное сопоставление** (undifferentiated mapping): Методика сопоставления, которая не различает разные типы соответствия и не указывает на различные степени эквивалентности.

См. термин «дифференцированное сопоставление».

3.89 **словарный контроль** (vocabulary control): Использование словаря для того, чтобы избежать многозначности и упорядочить форму представления терминов, а также ограничить число понятий и терминов, предназначенных для индексирования.

4 Символы, сокращения и условные обозначения

В настоящем стандарте применяются символы, сокращения и условные обозначения по ГОСТ Р 7.0.91—2015, раздел 3. Дополнительно используются метки и символы, приведенные в ГОСТ 7.75 и в таблице 1 настоящего стандарта.

В некоторых примерах ***метки, такие как «СЛ1» и «СЛ2», обозначают «Словарь 1», «Словарь 2» и т. п.***

Чтобы указать направление соответствия, одним из вариантов является использование стрелки. В качестве альтернативы, на языках, которые обычно читаются слева направо, понятие исходного словаря следует отражать слева, за ним давать соответствующую метку (метки) из таблицы 1, а за ней — понятие (понятия) целевого словаря.

Таблица 1 — Дополнительные сокращения и символы, используемые для обозначения соответствий

Символ	Значение
ЭК	Эквивалентность Термин, следующий за этим обозначением, является предпочтительным термином в целевом словаре, который наиболее близок по значению к предпочтительному термину исходного словаря
=	Знак равенства Этот символ (Unicode U + 003D) следует использовать вместе с аббревиатурой типа соответствия, для указания, что соответствие является точным. В частности, «=» означает точную эквивалентность
~	Тильда Этот символ (Unicode U + 007E) следует использовать вместе с аббревиатурой соответствия для указания, что соответствие является неточным. В частности, «~» означает неточную эквивалентность
ШС	Широкое соответствие Термин, следующий за этим обозначением, представляет понятие, имеющее более широкое значение
УС	Узкое соответствие Термин, следующий за этим обозначением, представляет понятие с более конкретным значением

Окончание таблицы 1

Символ	Значение
АС	Ассоциативное соответствие Термин, следующий за этим обозначением, представляет ассоциированное понятие, но не является синонимом, квазисинонимом, вышестоящим (более широким) или нижестоящим (узким) термином
|	Вертикальная черта Этот символ (Unicode U + 007C) разделяет два или более предпочтительных терминов целевого словаря, значения которых наилучшим образом охватывают значение более широкого предпочтительного термина в исходном словаре. Каждое из целевых понятий представляет собой часть объема исходного понятия. При преобразовании индексных терминов все предпочтительные термины целевого словаря должны применяться к записи, проиндексированной с помощью термина исходного словаря. При преобразовании поисковых предписаний предпочтительные термины целевого словаря должны быть объединены логическим оператором ИЛИ (OR). ***Пример — В исходном словаре — сельскохозяйственные животные, в целевом словаре — овцы | крупный рогатый скот | свиньи | птица***
+	Плюс Этот символ (Unicode U + 002B) ставят между двумя или более предпочтительными терминами из целевого словаря, которые используются совместно для представления составного понятия в исходном словаре. Каждое из целевых понятий представляет собой аспект исходного понятия. При преобразовании индексных терминов все предпочтительные термины целевого словаря должны применяться к записи, проиндексированной с помощью термина исходного словаря. При преобразовании поисковых предписаний предпочтительные термины целевого словаря должны быть объединены логическим оператором И (AND). ***Пример — В исходном словаре — женщины-руководители, в целевом словаре — женщины + руководство***

5 Задачи и идентификация

5.1 Задачи совместимости и соответствия

При поиске информации главная цель взаимодействия между словарями состоит в том, чтобы выражение, сформулированное с использованием одного словаря, преобразовать (или дополнить) в соответствующее выражение на основе одного или нескольких других словарей, независимо от того, используют ли эти словари один или разные естественные языки. Рассматриваемое выражение может быть либо поисковым запросом, либо частью метаданных, связанных с документом. В обоих случаях поиск соответствий является ключевым шагом. Если каждое из понятий Словаря А сопоставлено с соответствующим понятием (понятиями) Словаря В, открывается возможность замены (или дополнения) терминов или идентификаторов, представляющих понятия в каждом из словарей. На рабочем уровне совместимость обеспечивается установлением взаимных понятийных соответствий, в частности эквивалентности, для которых указания и рекомендации приведены в разделах 7—13.

Примечание — Другим аспектом установления совместимости является дополнение словарных инструментов такими операциями, как объединение нескольких словарей или использование частей одного для расширения другого. Данный стандарт не должен рассматриваться как ограничение новых форм взаимодействия, которые могут возникнуть в дальнейшем.

5.2 Идентификация элементов соответствия

Соответствия всех типов словарей показывают взаимосвязи между понятиями, хотя понятия обычно называют «классами» в контексте классификационной системы и «категориями» в случае таксономии. В настоящем стандарте термин «понятие» используется в широком смысле.

В таблице 2 приведены основные элементы, которые используются для представления понятий в разных типах словарей. Утверждения о соответствии должны использовать эти элементы, когда они предназначены для чтения людьми. Но когда соответствия предназначены для использования компьютером, понятия должны быть представлены уникальными постоянными идентификаторами.

Таблица 2 — Удобочитаемые элементы, используемые для представления понятий в утверждениях о соответствии

Тип словаря	Вид представления понятия
Тезаурус	Дескрипторы
Классификационная система	Классификационные коды
Таксономия	Метки или коды категорий (см. примечание)
Словарь предметных рубрик	Рубрики
Нормативный файл имен	Имена
Онтология	Метки
Терминосистема	Термины и другие виды обозначений
Примечание — Разные таксономии используют различные стили представления понятий. Если у понятий нет кодов или уникальных идентификаторов, а метки категорий не являются уникальными, обычно необходимо выписать весь иерархический путь, чтобы однозначно указать заданное понятие.	

6 Структурные модели соответствия между словарями

6.1 Общие положения

В этом разделе рассматриваются общие модели управления эквивалентностью и другими соответствиями. Три основные модели описаны в 6.2, 6.3 и 6.4, а альтернативный подход — в 6.5. Управление данными соответствий рассмотрено в разделе 15.

6.2 Модель 1: единая структура

В модели единой структуры все участвующие словари имеют одинаковую структуру иерархических и ассоциативных отношений между понятиями. Структура может быть представлена или выражена любым количеством различных языков, обозначений или систем кодирования. *Модель данных из ГОСТ Р 7.0.91—2015, раздел 15, иллюстрирует структурное единство на примере симметричного многоязычного тезауруса, рассматриваемого в ГОСТ 7.24.* Структурная простота этой модели делает возможным и желательным управление всеми понятиями, терминами, классификационными кодами, рубриками и отношениями между ними в рамках одной системы.

Примечание — Модель 1 обычно не предполагает соответствия, но включена сюда для полноты и сравнения с другими моделями. Некоторые приложения требуют применения Модели 1 в сочетании с другими моделями.

6.3 Модель 2: прямая связь

Модель прямой связи соответствует связям между двумя или более словарями, которые не разделяют одну и ту же структуру. Помимо различий в объеме, языке и структуре, словари могут принадлежать к разным типам (классификационные системы, нормативные файлы имен и т. д.), а также включать один или несколько тезаурусов. Между понятиями каждых двух словарей устанавливаются прямые соответствия. Эта модель может быть распространена на любое количество словарей путем установления прямых соответствий из одного словаря в другой. Каждый из словарей может затем использоваться для поиска в любой коллекции, индексированной по любому другому словарю.

На рисунке 1 показаны соответствия, необходимые для работы с четырьмя словарями с использованием модели прямой связи. Квадраты представляют понятия в четырех словарях соответственно. Соответствия представлены стрелками между понятиями. У каждой стрелки есть направление. Для простоты на рисунке 1 все соответствия показаны в виде двунаправленных стрелок, указывающих, что соответствия предназначены для работы в обоих направлениях. Чтобы описать это более точно, каждая двунаправленная стрелка представляет собой пару соответствий, по одному в каждом направлении. Таким образом, на рисунке 1 показано в общей сложности 12 наборов соответствий, представленных в шести парах.

Чтобы сократить объем соответствий, иногда устанавливаются соответствия только в одном направлении, что позволяет преобразовывать индексные или поисковые термины только в одном направлении. Соответствия в одном направлении будут представлены однонаправленными стрелками в той же базовой модели.

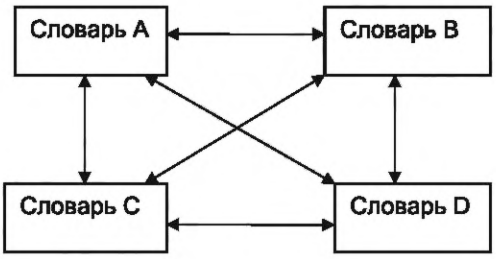

Рисунок 1 — Модель 2 (прямая связь) применительно к четырем словарям

6.4 Модель 3: централизованная структура

Вместо того, чтобы устанавливать прямые соответствия между каждой парой словарей, часто удобно обозначить один словарь как «хаб» или всеобъемлющую структуру, в которую отображается каждый из других словарей. На рисунке 2 показана Модель 3, в которой словарь B служит центром («хабом»), а остальные — «спутниками».

При применении этой модели каждое понятие в центральном словаре должно быть сопоставлено с соответствующим понятием (понятиями) в других словарях, и наоборот. В том же стиле, что и для Модели 2, двунаправленные стрелки указывают на соответствия в обоих направлениях.

Эта модель позволяет использовать центральный словарь для поиска любых ресурсов, проиндексированных с помощью любого из словарей-спутников: либо путем преобразования поисковых запросов, либо путем преобразования индексных терминов. Аналогично, индексный термин или поисковый термин, взятый из любого спутникового словаря, может быть преобразован в соответствующий термин (термины) в центральном словаре. Третья возможность состоит в том, чтобы применить два преобразования последовательно, например, из Словаря A в Словарь B, а затем из Словаря B в Словарь D. Когда оба этапа включают точную простую эквивалентность (см. раздел 8 и 11.2), последовательное соответствие дает вполне приемлемые результаты. Но если первый шаг включает в себя сложную эквивалентность или если два неточных соответствия применяются последовательно, качество результата непредсказуемо.

Для достижения хорошего качества соответствий для всего объема словарей-спутников центральный словарь должен включать в себя все понятия, присутствующие в спутниках. Когда среди отображаемых словарей нет словаря достаточного объема и специфики, может потребоваться создать новый центральный словарь для удовлетворения этой потребности. В качестве альтернативы, один из исходных словарей может служить в качестве центрального, если он расширен так, чтобы соответствовать объему и глубине всех остальных словарей.

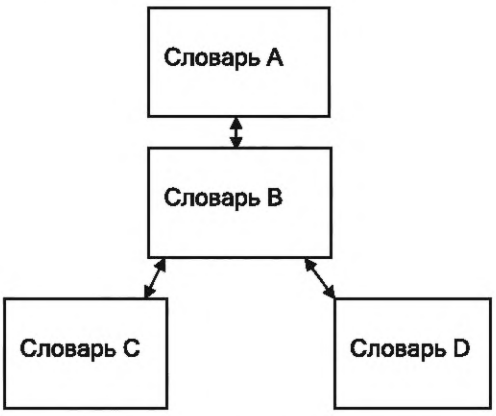

Рисунок 2 — Модель 3 (централизованная структура) применительно к четырем словарям

Для некоторых приложений двусторонние соответствия не нужны. На рисунке 3 показаны два варианта Модели 3, в которых соответствия применяются только в одном направлении.

- вариант 1 позволит пользователю Словаря Q искать ресурсы, проиндексированные с помощью Словарей P, R или S. В качестве альтернативы, если соответствия в нем используются для преобразования метаданных ресурсов, проиндексированных с помощью Словаря Q, пользователи Словарей P, R и S будут иметь возможность искать эти ресурсы;
- вариант 2 позволит пользователям Словарей W, Y и Z искать ресурсы, проиндексированные с помощью Словаря X. В качестве альтернативы, если соответствия в нем используются для преобразования метаданных ресурсов, проиндексированных с помощью Словарей W, Y или Z, пользователи Словаря X будут иметь возможность искать эти ресурсы.

И в том, и в другом случае невозможно преобразовать термины одного спутникового словаря в термины другого спутникового словаря.

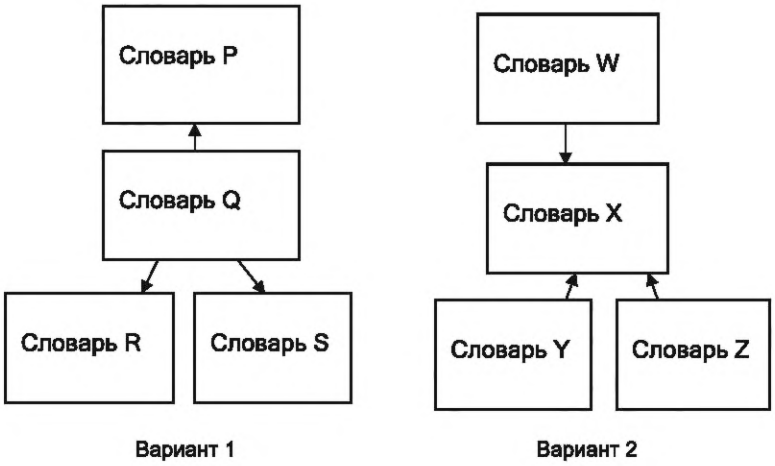

Рисунок 3 — Два варианта Модели 3 с односторонним соответствием

6.5 Селективное соответствие

Модели 1, 2 и 3 требуют значительной работы по разработке и ведению словарных соответствий. Однако для некоторых приложений нет необходимости отображать словари всесторонне. Одна из альтернатив состоит в том, чтобы установить соответствия только для тех понятий, которые использовались или могут использоваться в рассматриваемом приложении. Можно выделить два случая:

a) Область пересечения словарей мала.

Если имеется мало понятий, общих для двух или более словарей, то можно и нужно устанавливать только ограниченное число соответствий, как на рисунке 4.

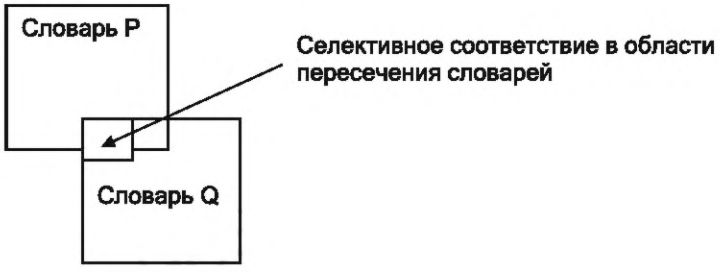

Рисунок 4 — Селективное соответствие при малой области пересечения словарей

b) Соответствие из указателя или каталога.

Способ, который часто оказывается удобным, состоит в том, чтобы установить соответствия рубрик (терминов или классификационных кодов), которые встречаются в указателе или каталоге коллекции (коллекций), включенных в сопоставление. При таком подходе указатель или каталог обрабатываются как подсловарь, полученный из полной версии словаря. Это уменьшает первоначальные усилия

по сопоставлению, но может усложнить работу по обновлению при внесении изменений в коллекцию. Рисунок 5 иллюстрирует этот подход применительно к каталогу.



Рисунок 5 — Соответствие единиц каталога и словаря

Когда каталог или указатель используется в качестве основы для сопоставления более чем с одним словарем, его можно рассматривать как центральный словарь из Модели 3. Обычно такие соответствия допускают преобразования только в одном направлении, от центрального словаря к словарям-спутникам.

6.6 Выбор модели соответствия

Три структурные модели были описаны отдельно, но в реальных приложениях часто встречаются их комбинации, и границы между ними размыты. В начале любой работы по сопоставлению словарей важно уточнить, какая модель или комбинация моделей будет использоваться, с особым вниманием к направлению соответствий.

При выборе вариантов следует учитывать следующие факторы:

a) Модель единой структуры (Модель 1) наиболее применима, когда готовится новый тезаурус для использования более чем на одном языке, как описано в ГОСТ Р 7.0.91—2015, раздел 15. Отсутствие индексированной коллекции облегчает предоставление каждому языку равного статуса;

b) Модели 2 и 3 подходят для выверки словарей, которые были независимо разработаны и/или уже применены к коллекциям. Модель 2 применима, в частности, когда есть только два или три таких словаря;

c) Модель 2 является более универсальной, чем Модель 3, если необходима прямая совместимость всех пар словарей, поскольку она исключает повторные операции отображения при поиске соответствий. Однако прямая совместимость между словарями-спутниками не всегда необходима. А затраты на техническое обслуживание Модели 2, как правило, выше;

d) для пары словарей может быть достигнуто наилучшее качество соответствия, когда целевой словарь имеет такую же специфичность и такую же широту охвата, что и исходный словарь (примеры соответствия в 16.2 иллюстрируют трудность, когда целевой словарь редок в требуемой предметной области);

e) как следствие пункта d), словарь, выбранный или созданный в качестве центрального, должен быть богат в требуемой предметной области (областях) на всех уровнях (то есть, он должен иметь как минимум одинаковую специфичность, а также такую же широту охвата, как и все словари-спутники вместе взятые);

f) если один словарь в группе сопоставляемых словарей имеет значительно меньшую специфичность, чем другие, то будет относительно мало эквивалентных соответствий. Более специфичные понятия в других словарях могут быть сопоставлены только иерархически или ассоциативно с более общими понятиями в первом словаре (см. разделы 7—10);

g) если многоязычный тезаурус включает в себя некоторые несимметричные структуры, может быть принят гибридный подход. Модель 1 должна использоваться везде, где это возможно, но Модель 2 может быть принята, когда оказывается невозможным установить полную эквивалентность между языками;

h) селективное соответствие на основе индекса или каталога, как описано в 6.5, имеет ограниченное применение. Эта модель может быть дешевле в реализации, но поддерживает соответствия для конкретного применения только в одном направлении. Поскольку в указателе отсутствуют некоторые понятия и отношения словаря, для которого он был разработан, семантическая структура словаря представляется неполно и, таким образом, она недостаточна для поддержки работы по установлению соответствий или для их применения.

7 Типы соответствий

7.1 Общие положения

Когда отношения устанавливаются через словари, их обычно называют соответствиями. В принципе, любой тип отношений может быть определен и применен как соответствие. На практике соответствия, которые могут оказаться полезными, определяются типами словарей. Следует рассмотреть три основных типа соответствий между тезаурусами: эквивалентные отношения, иерархические отношения и ассоциативные отношения (в точности аналогичные отношениям, используемым внутри любого тезауруса). Из них эквивалентность является наиболее часто используемым типом. Дополнительные типы отношений могут быть рассмотрены, если задействован другой тип словаря, особенно онтология.

Три основных типа соответствий и их подтипы описаны в разделах 8—10. В разделе 12 показано, как они используются при соответствии между тезаурусами. В разделе 13 рассматривается более сложный случай, когда одним из отображаемых словарей является предкоординированная система, например классификационная система или словарь предметных рубрик.

8 Эквивалентные соответствия

8.1 Общие положения

Эквивалентность устанавливается, когда в двух или более разных словарях обнаруживаются совпадающие понятия. В отличие от установления эквивалентности между двумя терминами в одном одноязычном тезаурусе, где один из них обозначен как предпочтительный термин (дескриптор), а другой — как непредпочтительный ***(аскриптор, см. ГОСТ Р 7.0.91—2015, раздел 8)***, в контексте межсловарной эквивалентности соответствие проводится между понятиями, и не выявляется разницы в статусе между понятиями или между дескрипторами или классификационными кодами, представляющими их. В записях о соответствии используется ***метка «ЭК».***

Как правило, словари включают в себя различные наборы понятий и развивают их до разных уровней специфики. Следовательно, может возникнуть несколько разных ситуаций эквивалентности с разными решениями, как описано в 8.2, 8.3 и разделе 11.

8.2 Простая эквивалентность

В идеальной ситуации целевой словарь содержит понятие, идентичное по объему понятию в исходном словаре. Между понятиями может быть немедленно установлено полное эквивалентное соответствие («один-к-одному»).

Пример —

Словарь 1	***Словарь 2***
mobile phones ***мобильные телефоны***	***cell phones*** ***сотовые телефоны***
maize ***маис***	***indian corn*** ***индийская кукуруза***

Соответствие обычно выражается в следующем формате: «мобильные телефоны ЭК сотовые телефоны».

Если рассматривается более двух словарей, в описании соответствия может потребоваться провести различия между ними. В разделе 16 показано, как это должно быть визуализировано для пользователей.

Очень часто дескрипторы для совпадающих понятий идентичны, и в таком случае уместно использовать такое описание соответствия, как «посудомоечные машины ЭК посудомоечные машины».

Обратное не всегда справедливо. Идентичные дескрипторы в двух разных словарях не должны считаться эквивалентами без проверки базовых понятий. Например, дескриптор «операции» может иметь разные значения в военном тезаурусе и в медицинском тезаурусе.

Даже когда контексты схожи, могут быть тонкие различия в значении. Например, понятие «учителя» в одном словаре может включать преподавательский состав как в университетах, так и в школах,

тогда как другой словарь может ограничивать сферу понятия «учителя» школьными учителями и предоставлять отдельное понятие «преподаватели» для преподавательского состава в университетах. Точно так же термин «public schools» имеет очень различное значение в контекстах Америки и Великобритании, потому что системы образования в этих странах организованы по-разному. В тех случаях, когда в разных словарях встречаются идентичные термины, соответствие эквивалентности следует устанавливать только в том случае, если лежащие в основе понятия оцениваются как эквивалентные.

Примечание — Дополнительные примеры эквивалентности, включая ситуации, в которых применяются разные степени эквивалентности, приведены в разделе 11.

8.3 Сложная эквивалентность

8.3.1 Общие положения

Сложное понятие, представленное в одном словаре одним термином, может быть представлено комбинацией двух или более понятий/терминов другого словаря.

Пример —

Словарь 1	***Словарь 2***
генетически модифицированная пшеница	***генетическая модификация*** ***пшеница***
внутренние водные пути	***каналы*** ***озера*** ***реки***
ископаемое топливо	***нефть*** ***природный газ*** ***торф*** ***уголь***

В таких случаях между понятиями может быть установлено сложное соответствие эквивалентности (также известное как эквивалентность «один-ко-многим»). Это соответствие обычно применяется только в одном направлении.

Следует различать два типа сложной эквивалентности, называемые пересекающаяся эквивалентность (эквивалентность пересечению) и кумулятивная эквивалентность (эквивалентность объединению).

8.3.2 Пересекающаяся сложная эквивалентность (ЭК +)

Соответствие типа «пересекающаяся сложная эквивалентность» называется так потому, что оно может быть представлено как пересечение двух или более множеств. Рисунок 6 иллюстрирует первый из следующих примеров.

Пример —

Словарь 1	***Словарь 2***
женщины-руководители	***женщины*** ***руководители***
генетически модифицированная пшеница	***генетическая модификация*** ***пшеница***
дети, подвергшиеся насилию	***дети*** ***насилие***
безопасность пассажиров железных дорог	***безопасность*** ***пассажиры*** ***железные дороги***

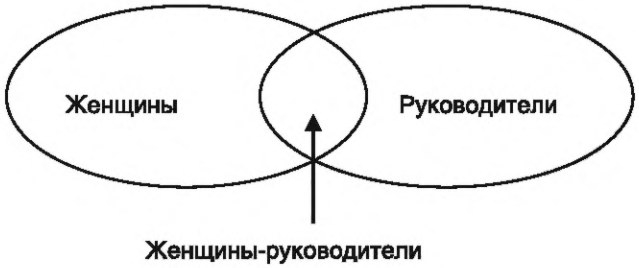

Рисунок 6 — Группа женщин, которые также являются руководителями, является подмножеством, показанным на пересечении двух больших множеств

Соответствие «пересекающаяся сложная эквивалентность» необратимо. Например, если «женщины» и «руководители» представлены в метаданных одного документа, эта комбинация не должна автоматически сопоставляться с «женщинами-руководителями», поскольку вместо этого документ может касаться отношений между руководителями-мужчинами и другими женщинами.

Использование некоторых примеров сложной эквивалентности иллюстрируется в 12.3.

Утверждение о пересекающейся сложной эквивалентности выражается в следующем формате: «женщины-руководители ЭК женщины + руководители».

Если соответствия будут использоваться только для преобразования поисковых предписаний, символ «+» в операторах соответствия может быть заменен булевым оператором «И». В контексте соответствия индексных терминов булевы операторы не применяются. Примеры использования соответствий в информационном поиске см. в 12.3.

8.3.3 Кумулятивная сложная эквивалентность (ЭК |)

Соответствие типа «кумулятивная сложная эквивалентность» устанавливается, когда сложное понятие в одном словаре соответствует сумме двух или более понятий в другом. Рисунок 7 показывает представление данного соответствия по теории множеств. Обратите внимание, что более простые понятия не обязательно накладываются друг на друга, но они составляют общее сложное понятие.

Пример —

Словарь 1	***Словарь 2***
внутренние водные пути	***каналы*** ***реки***
трикотаж	***носки*** ***чулки***
древесные растения	***деревья*** ***кустарник***
ископаемое топливо	***нефть*** ***природный газ*** ***торф*** ***уголь***

Соответствие «кумулятивная сложная эквивалентность» не должно использоваться в обратном направлении без проверки того, будет ли более подходящим другое соответствие. Например, в Словаре 1, помимо понятия «внутренние водные пути», могут быть представлены более узкие понятия «реки» и «каналы». В этом случае «реки» и «каналы» из Словаря 2 должны быть сопоставлены с их простыми эквивалентами, а не с более широким понятием.

Использование некоторых примеров сложной эквивалентности иллюстрируется в 12.3.



Рисунок 7 — Множество внутренних водных путей — это надмножество, состоящее из множества рек и множества каналов

Примечание — Этот пример работает в контексте регионов, где судоходны все реки, и нет озер или других водоемов, используемых в качестве внутренних водных путей. См. 11.4 с) для получения указаний относительно ситуации, когда более узкие понятия не составляют целое сложное понятие.

Утверждение о кумулятивной сложной эквивалентности следует выражать в следующей форме: «внутренние водные пути ЭК реки | каналы».

Если соответствия будут использоваться только для преобразования поисковых предписаний, символ «|» в операторах соответствия может быть заменен булевым оператором «ИЛИ». В контексте соответствия индексных терминов булевы операторы не применяются.

8.3.4 Сложная эквивалентность, включающая комбинацию целевых словарей

Особый случай сложной эквивалентности может возникнуть, когда целевой словарь неадекватен для передачи понятий в исходном словаре, если не сопровождается «словарем квалификаторов», т. е. словарем, составленным из понятий, которые могут использоваться для квалификации понятия в основном целевом словаре. В этом случае одно понятие исходного словаря может быть сопоставлено комбинации, в которой одно понятие извлекается из основного целевого словаря, а другое (другие) выбирается из словаря (словарей) квалификаторов (реляторов).

Пример —

Исходный словарь языков включает в себя полный список основных языков, а также региональных диалектов и других подъязыков. В этом словаре английский, на котором говорят в Великобритании, обозначен дескриптором «en-GB», в то время как тот, на котором говорят в Соединенных Штатах Америки, имеет дескриптор «en-US».

Для конкретного применения требуются соответствия для целевого словаря (VocL), который перечисляет основные языки, но не содержит положений для диалектов. Он сопровождается еще одним целевым словарем (VocG) с полным перечнем стран и регионов. Соответствия сложной эквивалентности могут быть установлены в следующем виде

en-GB ЭК VocL: английский + VocG: Великобритания

en-US ЭК VocL: английский + VocG: США.

Комбинации, описанные выше, следует использовать с осторожностью, поскольку они могут иметь ограниченную силу за пределами конкретного применения, для которого они предназначены. Например, индексный термин «Соединенное Королевство» обычно используется только для документов об этой стране и может ввести в заблуждение, если он применяется к ситуациям, которые не включают Великобританию.

9 Иерархические соответствия

Иерархическое соответствие может быть установлено между понятиями, когда одно из них явно шире другого. Так же, как и иерархические отношения в пределах одного тезауруса, ***описанные в ГОСТ Р 7.0.91—2015, раздел 11***, отношения между понятиями могут быть родо-видовыми или инстанциональными (элемент — множество). В отдельных случаях, ***описанных в ГОСТ Р 7.0.91—2015, раздел 11***, связи часть-целое также могут служить основанием для введения иерархических отношений.

Пример —

Словарь 1	***Словарь 2***
школьный персонал	***учителя***
дороги	***улицы***
международные организации	***Организация Объединенных Наций***
армии	***батальоны***
Соединенное Королевство	***Англия***

Иерархическое соответствие от более узкого понятия к более широкому понятию должно быть выражено в следующем формате: «улицы ШС дороги». Обратное иерархическое соответствие, от более широкого к более узкому понятию, должно быть выражено в следующем формате: «дороги УС улицы».

Примечание — ***Метки ШС и УС***, сокращенно обозначающие «широкое соответствие» и «узкое соответствие», аналогичны ***меткам «в» (выше) и «н» (ниже)***, используемым в пределах одного тезауруса. Использование различающихся меток позволяет избежать путаницы, когда визуализация включает как межсловарные, так и внутрисловарные соответствия.

Имеется возможность формально различать родо-видовые, инстанциональные и партитивные иерархические соответствия аналогично родо-видовым, инстанциональным и партитивным иерархическим отношениям в пределах одного тезауруса. Вместо меток BTG/NTG, BTI/NTI и BTP/NTP, описанных в ГОСТ Р 7.0.91, следует использовать метки BMG/NMG, BMI/NMI и BMP/NMP, ***которым в русском языке соответствуют метки ШСР/УСВ, ШСМ/УСЭ, ШСЦ/УСЧ соответственно***, как показано в таблице 3.

Таблица 3 — Утверждения о соответствии, которые различают подтипы иерархических соответствий

Подтип соответствия	Пример утверждения о соответствии	Обратное соответствие
Родо-видовое	крысы ШСР грызуны	грызуны УСВ крысы
Инстанциональное	Париж ШСМ столицы	столицы УСЭ Париж
Партитивное	пальцы ШСЦ руки	руки УСЧ пальцы

10 Ассоциативные соответствия

Ассоциативное соответствие может устанавливаться между понятиями, когда они не соответствуют требованиям эквивалентного или иерархического соответствия, но семантически связаны до такой степени, что документы, индексированные одним понятием, могут быть релевантны при поиске другого.

Разделительная линия между ассоциативным соответствием и неточной эквивалентностью (см. 11.3) является плохо определенной и субъективной, но может быть определена прагматически в соответствии с контекстом, в котором должны использоваться соответствия (принимая во внимание, например, интересы пользователей, объем релевантных ресурсов для поиска и способность систем поиска и сопоставления проводить четкие различия).

Пример —

Словарь 1	***Словарь 2***
дисциплина	***наказание***
электронное обучение	***дистанционное обучение***
письменный перевод	***устный перевод***

Ассоциативное соответствие выражается следующим способом: «дисциплина АС наказание». Ассоциативное соответствие действует в обе стороны, например, в данном случае справедливо «наказание АС дисциплина».

Примечание — ***Метка АС***, сокращение от «ассоциативное соответствие», аналогична ***метке «а» (ассоциация)***, используемой в пределах одного тезауруса. Использование различающихся меток позволяет избежать путаницы, когда визуализация включает как межсловарные, так и внутрисловарные соответствия.

11 Точная, неточная и частичная эквивалентность

11.1 Общие положения

Поскольку качество конкретного соответствия часто не является идеальным, можно к эквивалентному соответствию дополнительно применить маркер, указывающий степень, в которой соответствие универсально применимо. Различают две степени эквивалентности: точную (см. 11.2) и неточную (см. 11.3), которые должны быть отмечены символами «=» и «~» соответственно. Различие должно применяться только к случаям простой эквивалентности.

В случаях сложной эквивалентности маркеры точности/неточности не будут содержать никакой дополнительной информации, потому что все такие соответствия в некоторой степени неточны. И наоборот, в случае иерархических соответствий, должны приниматься только те соответствия, которые являются бесспорно «точными». Другие варианты более уместно характеризовать как ассоциативные соответствия или неточные эквиваленты. Наконец, в случае ассоциативных соответствий, они охватывают так много различных возможных типов отношений, что не имеет смысла обозначать их как точные или неточные.

11.2 Точная эквивалентность

Соответствие эквивалентности двух понятий, найденных в разных словарях, называется точным, когда понятия могут использоваться взаимозаменяемо во всех приложениях, которые могут быть предусмотрены для понятий. Если соответствия предназначены для использования только в одном или небольшом количестве контекстов, каждый из этих контекстов должен быть рассмотрен до определения эквивалентности как точной. Если соответствия применяются более широко, то решение должно приниматься по широкому спектру приложений для информационного поиска.

Для обозначения точной эквивалентности в утверждениях о соответствии должна использоваться ***метка «=ЭК».***

Пример —

Словарь 1	***Метка***	***Словарь 2***
коровье бешенство	***=ЭК***	***губчатая энцефалопатия крупного рогатого скота***
ангорские козы	***=ЭК***	***козы ангорские***

Точные соответствия эквивалентности по определению обратимы; другими словами, они могут применяться как двусторонние соответствия. При условии, что они были созданы корректно, они также могут без риска применяться в последовательных цепочках соответствия (см. 6.4).

11.3 Неточная эквивалентность

Иногда наиболее близко совпадающие понятия в двух или более словарях не совсем одинаковы. Проблема становится особенно острой, когда словари созданы в разных культурных сообществах ***(см. рассмотрение вопроса и примеры в ГОСТ Р 7.0.91—2015, раздел 10).*** Часто возникают следующие случаи:

- Понятия могут быть эквивалентны только в некоторых контекстах;
- Понятия могут иметь перекрывающиеся лексические области или небольшие различия в коннотации;
- При соответствии между классификационной системой и тезаурусом обычно находят класс с наименованием, которое соответствует дескриптору, но проверка показывает, что эти два значения не совсем эквивалентны. Причины этого рассматриваются в разделе 13 и 17.2.

Пример —

Словарь 1	*Словарь 2*
организационная структура	*структура управления*
децентрализованные органы власти	*децентрализация*
газоны	*дерн*
стулья	*сидения*
астрономия (в тезаурусе, как предпочтительный термин с «астрофизикой» в качестве ассоциативного термина)	*астрономия (в классификационной системе, как наименование широкого класса, который включает в себя астрофизику)*
женщины-руководители	*женщины-менеджеры*
плодоводство	*садоводство*

Неточное соответствие эквивалентности выражается следующим образом: «газоны ~ЭК дерн».

Неточная эквивалентность может иметь место как в случае сложной, так и простой эквивалентности, и должна быть выражена соответственно.

Когда возникает неточная эквивалентность между пересекающимися понятиями, может возникнуть вопрос, насколько велика область перекрытия. Этот вопрос важен в приложениях, где пользователю необходимо оценить вероятную выгоду от соответствия, чтобы получить более релевантную информацию, без большого числа нерелевантных документов. Оценка степени перекрытия также может помочь с ранжированием результатов поиска. Для таких приложений может оказаться полезным охарактеризовать неточную эквивалентность как «значительное пересечение» или «незначительное пересечение».

11.4 Частичная эквивалентность

Иногда единственное различие между родственными понятиями в разных словарях заключается в том, что одно немного шире другого. Иными словами, одно из понятий имеет объем, лишь частично эквивалентный другому.

Пример —

Словарь 1	*Словарь 2*
монархи	*короли*
контейнерные растения	*горшечные растения*
воздушные суда	*самолеты*
школьные помещения	*школьные здания*

Не существует уникального способа определения частичной эквивалентности. Вместо этого необходимо делать сложный, довольно субъективный выбор между следующими вариантами:

a) установить иерархическое соответствие (см. раздел 9), например, «монархи УС короли»;

b) установить соответствие эквивалентности, отметив его как неточное, если используется такое различие, например, «контейнерные растения ~ЭК горшечные растения». Эту опцию следует выбирать только в том случае, если иерархические соответствия недоступны в этом приложении, и опция «c» неприменима;

c) установить кумулятивную сложную эквивалентность. Это возможно только тогда, когда Словарь 2 также содержит понятие (понятия), содержащее недостающую часть понятия в Словаре 1. Например, «воздушные суда ЭК самолеты | вертолеты». Если считается, что важная часть более широкого понятия (например, воздушные шары, дирижабли и другие воздушные суда), скорее всего, будет упущена, следует установить два или более отдельных иерархических соответствия, а не одно сложное эквивалентное соответствие (например, «воздушные суда УС самолеты»; «воздушные суда УС вертолеты»).

12 Использование соответствий в информационном поиске

12.1 Общие положения

В контексте поиска информации существует два основных этапа, на которых могут использоваться соответствия между контролируемыми словарями: (a) как часть процесса индексирования или (b) во время поиска.

Примечание — Это не единственные ситуации, в которых могут использоваться соответствия, и нижеследующие рекомендации не следует интерпретировать как исключающие другие варианты применения.

a) Когда соответствия используют в процессе индексирования, индексные термины в метаданных документов, индексируемых с помощью Словаря A, преобразуются (или дополняются) в соответствующие термины из Словаря B. Это можно делать регулярно при первоначальном индексировании, или как пакетное преобразование всей коллекции, дополняемое небольшими обновлениями, каждый раз при обновлении словарей и/или при добавлении в коллекцию новых документов. Коллекцию документов затем можно просматривать или искать в ней информацию с использованием Словаря B без необходимости дополнительного сопоставления словарей.

b) Когда соответствия применяют к поисковым терминам, исходные метаданные не изменяются. Для достижения той же возможности (использования Словаря B для поиска в коллекции, индексируемой с помощью Словаря A), необходимо установить исходный и целевой словари для соответствий в обратном направлении. Это позволяет преобразовывать поисковые запросы, содержащие термины из Словаря B, в соответствующие термины из Словаря A. Таким образом, соответствия включаются в процесс поиска и должны применяться каждый раз, когда поиск выполняется с использованием Словаря B.

12.2 Участие человека

В некоторых приложениях соответствия применяются автоматически, но в других случаях процесс контролируется специалистом, который проверяет пригодность каждого соответствия. Это особенно полезно для соответствий, отличных от точной простой эквивалентности. В тех случаях, когда нет точного эквивалента, индексатор делает выбор между любыми неточными и/или иерархическими соответствиями, выбирая то, которое лучше всего подходит для рассматриваемого документа. Аналогичным образом, пользователь может выбрать соответствие, наиболее близкое к его собственной информационной потребности.

Пример — Пользователь ищет информацию о «защите детей», что является дескриптором в Словаре A. Но в Словаре B этот термин отсутствует, и вместо него пользователю предлагается неточное соответствие — «охрана детей» и/или ассоциативное соответствие — «уход за детьми». Пользователь может выбрать один или оба эти термина, в зависимости от того, какой из них наиболее близок к его потребности.

Участие человека следует использовать везде, где это возможно, для достижения более высокой эффективности поиска и во избежание ложных выводов из-за неточных эквивалентностей.

12.3 Примеры реализации соответствий

При простейшем подходе к сопоставлению словарей цель состоит только в том, чтобы указать пары понятий в разных словарях, которые имеют некоторое неопределенное соответствие друг другу. Однако точность и полнота информационного поиска могут быть значительно повышены, если сопоставления дифференцируются с использованием всех типов соответствия, описанных в разделах 7—10, и если степени эквивалентности отмечены, как это рекомендуется в разделе 11.

Чтобы проиллюстрировать силу дифференциации, в таблице 4 приведены примеры того, как различные типы соответствий работают на практике. Для простоты предполагается, что во всех случаях используемые словари являются тезаурусами (раздел 13 показывает более сложные примеры, в которых один из словарей представляет собой моноиерархическую схему, такую как классификационная система).

В первой колонке таблицы 4 приведен пример каждого из основных типов соответствий, описанных в разделах 7—11.

Во второй колонке первое предложение для каждого примера говорит о том, как это должно быть реализовано, если соответствия используются для преобразования из Словаря A в Словарь B индекс-

ных терминов. Предполагается автоматическое преобразование без какого-либо посредничества со стороны человека. Дополнительно комментируется вероятная эффективность сопоставленных индексных терминов, когда для поиска используется Словарь В. Эффективность сопоставленных терминов сравнивается с возможной эффективностью, если бы те же документы были проиндексированы непосредственно с помощью Словаря В, а не косвенно с помощью соответствия из Словаря А. Оценка эффективности неизбежно является несколько гипотетической, поскольку индексные термины используются для множества различных поисков, где характеристики точности/полноты будут различны.

В третьей колонке первое предложение для каждого примера говорит о том, как реализуется поиск, если соответствия используются для преобразования поисковых терминов из Словаря А в Словарь В. Предполагается автоматическое преобразование без какого-либо посредничества со стороны человека. Дополнительно приводится комментарий о вероятной эффективности сопоставленного поискового выражения. Вероятный эффект полноты и точности сравнивается с возможной эффективностью, если бы коллекция документов была проиндексирована с помощью Словаря В.

Таблица 4 — Сравнение реализации соответствий при преобразовании индексных терминов и поисковых терминов

Тип соответствия (пример)	Применение к индексным терминам	Применение к поисковым терминам
Точная простая эквивалентность: коровье бешенство =ЭК губчатая энцефалопатия крупного рогатого скота	Индексный термин «коровье бешенство» преобразован в «губчатая энцефалопатия крупного рогатого скота». Никаких отрицательных эффектов не должно быть, если контролируемый словарь используется корректно	Поиск «коровье бешенство» заменен на поиск «губчатая энцефалопатия крупного рогатого скота». Как точность, так и полнота, такие же, как если бы коллекция была проиндексирована целевым словарем
Пересекающаяся сложная эквивалентность: женщины-руководители ЭК женщины + руководители	Оба сопоставленных широких термина добавлены в метаданные. Это способствует повышению полноты при всех поисках по терминам «женщины» и «руководители». Но пользователи, ищущие именно женщин-руководителей, будут в затруднении при выделении релевантных документов среди документов, относящихся к руководителям обоих полов и к женщинам на любых должностях	Поисковое предписание стало: «женщины И руководители». Полнота не изменилась. Точность уменьшилась. Этот поиск выдаст документы о руководителях обоих полов и о женщинах; например, о руководителях-мужчинах, которые оскорбляют женщин
Кумулятивная сложная эквивалентность: внутренние водные пути ЭК реки | каналы	Оба сопоставленных узких термина добавлены в метаданные. Если Словарь А также содержит термины «реки» и «каналы», то возможно снижение точности (поскольку поиск «каналов» может привести к выдаче документов, индексированных термином «внутренние водные пути», но рассматривающих только реки)	Поисковое предписание стало: «реки ИЛИ каналы». Негативного воздействия на полноту нет, но точность может снизиться, если пользователь хочет провести поиск только на верхнем уровне детализации
Неточная простая эквивалентность: горшечные растения ~ЭК комнатные растения	Индексный термин «горшечные растения» заменен на «комнатные растения». Для поиска комнатных растений точность будет меньше, поскольку будут выданы документы о горшечных растениях на открытом воздухе. Полнота также может быть снижена, если Словарь А не включает понятие «комнатные растения», и документы о негоршечных комнатных растениях не будут индексированы ни одним из этих терминов	Поисковый термин «горшечные растения» заменен на «комнатные растения». Точность уменьшилась, поскольку выдаются отдельные документы о комнатных растениях в контейнерах, а не в горшках

Окончание таблицы 4

Тип соответствия (пример)	Применение к индексным терминам	Применение к поисковым терминам
Иерархическое соответствие (от узкого к широкому): улицы ШС дороги	Индексный термин «улицы» заменен на «дороги». Если контролируемый словарь используется эффективно, то негативные эффекты будут незначительными	Поисковый термин «улицы» заменен на «дороги». Полнота может незначительно возрасти, а точность снизиться, поскольку запрос расширился
Иерархическое соответствие (от широкого к узкому): дороги УС улицы	Индексный термин «дороги» заменен на «улицы». Это окажет негативное влияние на точность поиска по «улицы», поскольку будут выданы документы о всех типах дорог (например, автострадах и проселочных дорогах), которые обычно не рассматриваются как улицы	Поисковый термин «дороги» заменен на «улицы». Точность не пострадает, но полнота уменьшится, поскольку будут выданы документы только об одном типе дорог
Ассоциативное соответствие: электронное обучение АС дистанционное обучение	Индексный термин «электронное обучение» заменен на «дистанционное обучение». В результате пользователи, ищущие дистанционное обучение, могут рассчитывать на получение релевантных документов вместе с некоторым количеством документов, рассматривающих другие приложения электронного обучения	Поисковый термин «электронное обучение» заменен на «дистанционное обучение». Ожидается негативное воздействие как на полноту, так и на точность

12.4 Выводы и рекомендации

Примеры в 12.3 показывают, что некоторые типы соответствий более полезны, чем другие, в зависимости от контекста применения и относительной специфичности словарей:

- точная простая эквивалентность является единственным типом, который не влияет на качество поиска и может быть применен обратимо в любом контексте, как для индексных терминов, так и для поисковых терминов;
- кумулятивная сложная эквивалентность также может быть очень эффективной при условии, что она применяется в заданном направлении, и сочетание более узких понятий действительно охватывает все сопоставляемое понятие;
- пересекающаяся сложная эквивалентность обычно снижает точность поиска;
- ни одна из форм сложной эквивалентности не может быть использована в обратном направлении. То есть совместное вхождение двух или более терминов в поисковое предписание или метаданные документа не должно без других проверок служить основанием для вывода об их эквивалентности одному комбинированному понятию. Однако отдельные термины можно использовать для установления иерархических соответствий в обратном направлении. Например, как «женщины» и «руководители», так и их комбинация предполагают иерархическое соответствие с более узким понятием «женщины-руководители». И наоборот, как «реки» и «каналы», так и их комбинация, предполагают иерархическое соответствие более широкому понятию «внутренние водные пути»;
- иерархические соответствия имеют тенденцию давать более приемлемые результаты в направлении от более узкого к более широкому понятию, чем в обратном направлении;
- в конкретных приложениях достигается бóльшая гибкость, когда сопоставление не ограничивается одним соответствием на понятие. Особенно это желательно тогда, когда понятие не имеет точного эквивалента. Предоставление неточных, более широких, более узких и ассоциативных соответствий помогает в выборе лучшего варианта для требуемого контекста.

В начале любой работы над соответствиями словарей необходимо выбрать:

- какую общую модель или комбинацию моделей использовать (см. раздел 6);
- насколько сильно дифференцировать соответствия в следующих отношениях:
- следует ли проводить различие между эквивалентностью и другими типами соответствия, такими как иерархическое и ассоциативное (см. разделы 7—10);

- применять ли сложные соответствия эквивалентности (см. 8.3);
- следует ли различать точную и неточную эквивалентность (см. раздел 11);
- разрешить ли указание более одного соответствия на понятие;
- применять ли соответствия в точке индексирования или в точке поиска;
- в каком направлении устанавливать соответствия, и требуются ли двусторонние соответствия;
- следует ли предусматривать участие человека в процессе конверсии терминов.

13 Управление предкоординацией

13.1 Общие положения

Некоторые типы словарей предусматривают предкоординированные понятия или рубрики, то есть сложные понятия, в которых два или более простых понятия объединены в одно. Иногда предкоординированные рубрики указаны явно.

Примеры явного указания —

a	***344.032 Закон о социальном обеспечении***	***по ДКД (Десятичная классификация Дьюи)***
b	***Книги — Африка — История***	***по LCSH (Предметные рубрики Библиотеки Конгресса)***
c	***Трубопроводы — Защитные покрытия***	***по ПР РНБ (Предметные рубрики Российской национальной библиотеки)***

В дополнение к явному указанию, некоторые словари включают в себя правила синтеза классов или рубрик, когда это необходимо, путем сочетания более простых элементов при индексировании.

Примеры синтезирования —

c	***373.3.016:51 (Преподавание математики в начальной школе)***	***Этот класс не представлен в УДК, а синтезирован из комбинации класса 373.3 (Начальная школа. Начальный уровень обучения) с классами 37.016 (Курс обучения) и 51 (Математика)***
d	***Пирамиды — Египет***	***Эта рубрика не представлена в LCSH, а синтезирована из перечисленных рубрик «Пирамиды» и «Египет»***

В классификационных системах предкоординированные понятия не всегда ясно отражены в наименовании класса. Например, в схеме, приведенной на рисунке 8, наименование класса «учреждения» представлено в трех разных иерархиях, относящихся к трем различным разделам. В рамках раздела «Образование» это наименование подразумевает, что подкласс E100 относится к образовательным учреждениям. В разделе «Оборона» то же самое наименование используется с целью указания, что D100 покрывает область оборонных учреждений.

E Образование	H Здоровье	D Оборона
E100 учреждения	H100 учреждения	D100 учреждения
E200 деятельность	H200 деятельность	D200 деятельность
E210 обучение *медицинское обучение, см. H210* *военная подготовка, см. D210*	H210 обучение H220 профилактическая медицина	D210 обучение D220 боевые действия
E211 подготовка учителей	H230 операции	D230 обеспечение тыла
и т. д.	и т. д.	и т. д.

Рисунок 8 — Фрагмент моноиерархической системы с классификационными кодами

Появление предкоординированных понятий, классов или рубрик создает дополнительную проблему для совместимости. Когда в двух словарях встречается одно и то же предкоординированное понятие, можно и нужно установить взаимное соответствие («один-к-одному»). Чаще всего выбор со-

ставляющих понятий для комбинирования варьирует от одного словаря к другому, и это приводит к необходимости соответствия «один-ко-многим».

Предкоординированные рубрики в стиле примеров *b)* и *d)*, указанных выше, встречаются только в словарях предметных рубрик. Способ их обработки описан в разделе 20.

Стиль предкоординации, приведенный на рисунке 8, встречается не только в классификационных системах, но и часто в любых системах с моноиерархической структурой, например, в системах управления документами и других системах регистрации, а также во многих таксономиях. Сопоставление тезауруса с этими системами описано в 13.2.

Классы классификационных систем обычно являются сложными по существу, как приведено в примере *c)* выше. Другие примеры такого рода описаны в 13.3.

13.2 Сопоставление тезауруса и системы с контекстно-зависимыми рубриками

13.2.1 Общие положения

При установлении соответствия между классами или категориями в моноиерархической схеме, класс/категорию необходимо рассматривать как предкоординированное понятие, значение которого может быть установлено путем проверки всех его вышестоящих и подчиненных классов, а также любых лексических примечаний, связанных с ним. Учитывать смысл только наименования класса недостаточно. Например, на рисунке 8 «учреждения» не отвечают требованиям класса E100, который применяется только к образовательным учреждениям.

При сопоставлении может быть использован любой тип соответствия — эквивалентность, иерархическое и ассоциативное соответствие (см. разделы 7—10). Если одно понятие в тезаурусе и класс в системе соответствуют друг другу по одному из этих типов, может быть установлено соответствие «один-к-одному».

Когда для данного класса нет соответствующего понятия, особенно если класс включает в себя предкоординацию, тогда класс может быть проанализирован на предмет возможности найти соответствие «один-ко-многим». Этот процесс иногда называют разложением класса. Так класс 373.3.016:51 из примера c) в 13.1 можно разложить на более простые классы — 373.3 (Начальная школа), 37.016 (специальный определитель понятия «курс обучения», от которого в комбинированном индексе используется только .016) и 51 (Математика); и затем найти соответствия для этих более простых классов в тезаурусе.

В утверждении о соответствии предкоординированное понятие должно быть однозначно представлено его обозначением или специально созданным идентификатором. Наименование без кода не должно использоваться для представления класса в утверждении о соответствии.

Чтобы проиллюстрировать этот раздел, использована структура, приведенная на рисунке 8, для предоставления примеров. Эта структура может быть частью очень простой классификационной системы, моноиерархической таксономии или схемы, используемой системой управления документами. Рекомендации в равной степени применимы ко всем этим и к любому другому словарю, следующему аналогичной предкоординированной моноиерархической модели. Как и в разделе 12, сопоставление в процессе индексирования будет рассматриваться отдельно от сопоставления при поиске.

13.2.2 Соответствия для преобразования записей индексирования/каталогизации, когда тезаурус является целевым словарем

Необходимо соблюдать следующие рекомендации:

a) необходимо найти соответствие для каждого класса классификационной системы. Примеры в таблице 5 основаны на классах, приведенных на рисунке 8 в иерархиях «Образование» и «Здоровье»;

b) для каждого класса следует искать наиболее близко совпадающее понятие или комбинацию понятий тезауруса, изучая дескрипторы, аскрипторы и лексические примечания по каждому предлагаемому понятию. В случаях сомнений следует также проверять вышестоящие и нижестоящие термины, а также способы использования дескрипторов при индексировании;

c) если найдено точное эквивалентное понятие тезауруса, обычно достаточно установить одно соответствие эквивалентности. Но если наилучшее доступное совпадение является неточным, могут быть установлены дополнительно более широкие, более узкие или ассоциативные соответствия для всех предлагаемых понятий, которые кажутся полезными в рассматриваемом контексте;

d) там, где установлена точная простая эквивалентность, соответствия могут использоваться для автоматического преобразования классификационных кодов метаданных в соответствующие индексные термины. В других случаях рекомендуется применять дополнительные проверки, включая участие человека в процессе (см. 12.2).

Таблица 5 — Образцы сопоставления классификационной системы тезаурусу

Класс	Соответствие в тезаурусе	Утверждение о соответствии
E	«Образование» — предпочтительный термин со множеством нижестоящих и ассоциативных терминов	«E ЭК образование» Может быть добавлен маркер = или ~ в зависимости от того, предполагает ли тезаурус понимание термина в том же смысле, что и класс E
H	Термины «здоровье» и «здоровье человека» отсутствуют, но найдены предпочтительные термины «общественное здравоохранение» и «здоровье животных»	«H ЭК общественное здравоохранение» Может быть добавлен маркер = или ~ в зависимости от того, предполагает ли тезаурус понимание термина в том же смысле, что и класс H. Альтернативным соответствием могло бы быть «E ЭК общественное здравоохранение | здоровье животных», но только в том случае, когда подклассы класса H включают вопросы здоровья животных наряду со здоровьем человека
E100	Найден дескриптор «учреждения» с нижестоящими терминами «образовательные учреждения», «больницы» и «клиники»	«E100 ЭК образовательные учреждения» Может быть добавлен маркер = или ~ в зависимости от того, предполагает ли тезаурус понимание термина в том же смысле, что и класс E100
H100	Найден дескриптор «учреждения» с нижестоящими терминами «больницы» и «клиники». Но отсутствуют понятия, отражающие оздоровительные центры и медицинские учреждения в целом, а также более специализированные учреждения, например зубоврачебные кабинеты и центры фитнеса	«H100 ЭК больницы | клиники» или «H100 УС больницы», а также «H100 УС клиники» Маркер ~ здесь был бы излишен, поскольку сложные эквивалентности всегда до некоторой степени неточны. Класс H100 далек от точного соответствия объединению больниц и клиник, поскольку они не исчерпывают весь объем понятия «медицинские учреждения». Второй вариант соответствия, который явно маркирует найденные термины как более узкие понятия, выражает ситуацию более ясно
E210	Найден дескриптор «обучение» с нижестоящими терминами «обучение первой помощи» и «обучение пожарной безопасности», а также с ассоциациями «образование» и «непрерывное повышение квалификации»	«E210 ~ЭК обучение» или «E210 ШС обучение» Выбор варианта зависит от того, предполагает ли E210 обучение деятельности только в сфере образования (тогда понятие «обучение» в тезаурусе предстает как более широкое понятие) или покрывает все виды обучения. В последнем случае более уместна неточная эквивалентность
H210	Найден дескриптор «обучение» с нижестоящими терминами «обучение первой помощи» и «обучение пожарной безопасности», а также с ассоциациями «образование» и «непрерывное повышение квалификации». Также присутствует предпочтительный термин «медицинское образование»	«H210 УС обучение первой помощи» «H210 ШС обучение» «H210 АС непрерывное повышение квалификации» «H210 АС медицинское образование» В этом случае ясно, что тезаурусное понятие «обучение» гораздо шире класса H210, который очевидным образом ограничен сферой здравоохранения. Однако ничего приблизительно эквивалентного H210 не найдено, и потому остается только один вопрос, насколько полно следует устанавливать возможные утверждения о соответствии
E211	Найдены дескрипторы «учителя» и «обучение»	«E211 ЭК учителя + обучение» Это соответствие довольно неоднозначно, поскольку ту же самую комбинацию можно использовать и для обучения учителями, и для обучения учителей. Однако при отсутствии в тезаурусе понятия «обучение учителей» данный вариант является наилучшей возможностью

Окончание таблицы 5

Класс	Соответствие в тезаурусе	Утверждение о соответствии
H220	Найдены дескрипторы «медицина» и «вакцинация»	«H220 УС вакцинация» или «H220 АС вакцинация» «H220 ШС медицина» Узкое отображение на «вакцинацию» выглядит более надежным, поскольку вакцинация всегда может рассматриваться как форма профилактической медицины. Но если тезаурусный термин использовался для ветеринарной вакцинации, выходя за рамки класса H в классификационной системе, то ассоциативное соответствие будет более адекватным. Предлагаемое широкое соответствие термину «медицина» также требует тщательной проверки, так как может быть, что это тезаурусное понятие прилагается только к учебной дисциплине, а не к врачебной практике
H230	В тезаурусе не найдено понятий, хоть отдаленно связанных с операциями (в медицинском смысле)	Тогда лучше совсем не устанавливать соответствий, чем вводить ложные связи

13.2.3 Соответствия для преобразования поисковых предписаний, когда тезаурус является целевым словарем

Рекомендации a)—c), а также примеры в 13.2.2 применяются также к соответствиям при преобразовании поисковых предписаний.

Когда подвергаются преобразованию поисковые выражения, соответствия сложной и простой эквивалентности могут применяться автоматически с небольшим ухудшением качества поиска. Там, где также присутствуют иерархические и ассоциативные соответствия, рекомендуется участие человека (например, выбор между доступными альтернативными соответствиями), см. 12.2.

13.2.4 Соответствия для преобразования записей индексирования/каталогизации, когда тезаурус является исходным словарем

Обычной целью классификации является присвоение документу единого кода, объединяющего основные моменты содержания так, чтобы описать их трактовку в документе. Это не должно делаться исключительно на основе индексных терминов, выделенных для посткоординатного использования, поскольку они могут быть скоординированы с неправильным синтаксисом. Например, если индексные термины, присвоенные документу, включают «женщин» и «руководителей», документ может касаться обращения руководителей с женщинами или исполнения женщинами руководящих функций, среди других вариантов, и каждый из этих вариантов потребовал бы другого кода. Следовательно, создание классификационного кода, полученного из совместного использования только индексных терминов, не рекомендуется.

13.2.5 Соответствия для преобразования поисковых предписаний, когда тезаурус является исходным словарем

В таблице 6 приведены примеры соответствий терминов тезауруса классификационной системе, представленной на рисунке 8 (где показаны фрагменты, а не вся система).

Применяются следующие рекомендации:

a) для каждого понятия тезауруса должно быть найдено, по крайней мере, одно соответствие;

b) дополнительные соответствия должны быть установлены во всех случаях, когда понятие находится в предкоординированном классе, который указан в моноиерархической системе;

c) когда установлено эквивалентное соответствие, обычно нет необходимости устанавливать соответствия с какими-либо подклассами рассматриваемого класса. Таким образом, в примере a) таблицы 6, если выбрано соответствие «образование ЭК E», нет необходимости устанавливать прямое соответствие «образования» и «E100» или любого другого класса в разделе «Образование» (эта рекомендация предполагает, что поисковая система может использовать структуру классификационной системы для расширения поиска определенного класса на любой из его подклассов);

d) в дополнение к соответствиям, установленным путем проверки классов, перечисленных в системе, следует рассмотреть классы и классификационные коды, синтезированные в соответствии с правилами системы. Одним из таковых является соответствие, показанное в примере f) таблицы 6, где конкретное соответствие может быть синтезировано для понятия тезауруса. Другим является расширение любого поиска для идентификации соответствующих строк, встроенных в синтезированные клас-

сификационные коды. Например, поиск «медицинской подготовки» может быть расширен до «H210», причем не по целому классификационному коду, а по фрагменту более длинного классификационного кода, такого как «S210(H210)». Чтобы повысить до предела полноту, поиск должен быть расширен данным образом, даже если может быть установлено соответствие точной простой эквивалентности, как в примерах a), d), e) и g). Однако расширение поиска для извлечения таких фрагментов из строк синтетического классификационного кода следует выполнять с особой тщательностью, поскольку правила синтеза и анализа часто намного сложнее, чем предполагает этот пример;

e) если запрос содержит более одного понятия тезауруса и если пользователь не может проверить результат соответствия, термины следует преобразовать отдельно, а не пытаться объединить в один синтезированный классификационный код. Например, если запрос включает термины «учителя» и «обучение», преобразование в «E211» будет неадекватным, поскольку неясно, ищет ли пользователь информацию об *обучении учителей* или *обучении у учителей*;

f) чтобы максимально увеличить точность и полноту, рекомендуется допустить участие человека (или какую-либо другую форму проверки) в процессе поиска (см. 12.2).

Таблица 6 — Простые соответствия тезауруса классификационной системе

Пример	Понятие тезауруса	Соответствие в классификационной системе	Утверждения о соответствии и комментарии
a	образование	E Образование	«образование ШС E» или «образование ЭК E» Выбор зависит от того, предназначено ли понятие тезауруса для использования в том же смысле, что и класс E. Это может быть установлено путем более тщательного изучения обоих словарей и, предпочтительно, также того, как они использовались при классифицировании/индексировании коллекций для поиска. Так как класс E очень широк и охватывает как учебные заведения, так и процесс обучения, иерархическое соответствие, вероятно, будет более подходящим. Но если выбрана эквивалентность, может быть применен маркер «~» или «=»
b	оборонный сектор	D Оборона	«оборонный сектор ШС D» или «оборонный сектор ЭК D» Выбор зависит от того, предназначен ли класс D для использования в том же смысле, что и понятие тезауруса. Ни одно из этих соответствий не будет уместным, если D имеет подкласс, который точно соответствует значению «оборонного сектора»
c	учреждения	E100 учреждения H100 учреждения D100 учреждения	«учреждения ЭК E100 | H100 | D100» Если в системе есть дополнительно ветви иерархии других секторов, таких как Транспорт, Спорт и т. д., и в каждом из них есть класс учреждений, эти классы также должны быть включены в соответствие. Альтернативный подход заключается в создании отдельных иерархических соответствий для всех соответствующих классов, например «учреждения УС E100», «учреждения УС H100» и т. д.
d	образовательные учреждения	E Образование E100 учреждения	«образовательные учреждения ЭК E100» При желании можно использовать маркер «~» или «=» после рассмотрения лексических примечаний, нижестоящих дескрипторов и другой контекстной информации в обоих словарях
e	школы	E110 школы	«школы ЭК E100» При желании можно использовать маркер «~» или «=» после рассмотрения лексических примечаний, нижестоящих дескрипторов и другой контекстной информации в обоих словарях

Окончание таблицы 6

Пример	Понятие тезауруса	Соответствие в классификационной системе	Утверждения о соответствии и комментарии
f	школы дайвинга	E110 школы S Спорт S100 учреждения S110 учебные центры S200 виды деятельности S245 плавание и дайвинг	«школы дайвинга ШС S110:245» Это соответствие использует действующее в рамках этой классификационной системы правило, позволяющее синтезировать классификационные коды для учреждений, специфичных для того или иного вида спорта, как показано выше. Подходящий для данного случая тип соответствия является иерархическим, так как синтезированный классификационный код включает в себя учебные центры для плавания, а также для дайвинга. Можно было бы рассмотреть соответствие с E110, но оно будет отклонено, потому что правила системы поясняют, что школы дайвинга не входят в этот класс
g	медицинская подготовка	H Здоровье H210 обучение	«медицинская подготовка ЭК H210» Это соответствие, вероятно, будет применяться при условии, что система также не включает предкоординированный класс, включающий какие-либо аспекты медицинской подготовки. Например, в иерархии «D Оборона» может существовать класс, такой как D210(H210), включенный как класс специальной медицинской подготовки в войсках. Приведенное выше соответствие в этом случае будет изменено следующим образом: «медицинская подготовка ЭК H210 | D210 (H210)»
h	обучение	E210 обучение E211 подготовка учителей H210 обучение D210 обучение S210 обучение	«обучение ЭК E210 | H210 | D210 | S210» или все следующие: «обучение УС E210»; «обучение УС H210»; «обучение УС D210»; «обучение УС S210» Нет необходимости включать какое-либо прямое соответствие к E211, поскольку E211 уже неявно включен в вышестоящий класс E210

13.3 Соответствия более сложных классов

Классификационные системы, особенно крупные, часто включают классы более сложные, чем примеры, показанные в 13.2. Для них при формулировке соответствия может потребоваться комбинация символов + и |.

Пример

	Найдено в классификационной системе	***Найдено в тезаурусе***	***Формулировка соответствия***
a	***T563 внутренний водный транспорт***	***реки каналы транспорт***	***T563 ЭК (реки | каналы) + транспорт***
b	***629.276, с наименованием «Аксессуары для обеспечения безопасности» в иерархии «Наземные моторные транспортные средства, велосипеды». В классе также есть примечание: «Включая подушки безопасности, бамперы, зеркала, ремни безопасности, дворники и омыватели ветрового стекла»***	***стеклоочистители подушки безопасности ремни безопасности удерживающие устройства (безопасность автомобиля) автотранспорт велосипеды***	***629.276 ЭК (моторные транспортные средства | велосипеды) + (стеклоочистители | подушки безопасности | ремни безопасности | удерживающие устройства (безопасность автомобиля))***

Пример b) иллюстрирует необходимость учета вышестоящих и подчиненных классов, а также сопровождающих примечаний при определении охвата рассматриваемого класса. Это также поднимает вопрос о том, насколько надежно формулировка такого сложного соответствия будет интерпретироваться в реальных поисковых ситуациях. Например, булевы операторы обычно не являются допустимыми или значимыми при применении к индексным терминам в метаданных. И не все поисковые системы будут надежно воспринимать вложенные скобки и/или проводить различие между теми скобками, которые обозначают реляторы терминов, и теми, которые образуют синтаксис поискового предписания.

Описания некоторых классов могут исключать определенные аспекты. Например, класс ДКД 331.8811 охватывает «профсоюзы в отраслях промышленности и профессиях, отличных от добывающей, обрабатывающей промышленности и строительства». Возникает вопрос о том, следует ли сопоставлять этот класс с такой комбинацией, как «профсоюзы НЕ (добыча ИЛИ производство, ИЛИ строительство)». Такое соответствие имеет очень ограниченную полезность. Если оно используется для преобразования метаданных каталогизации в индексные термины из тезауруса, то, безусловно, можно выбрать такой термин тезауруса, как «профсоюзы» или «профессиональные союзы», если он присутствует в тезаурусе. Но современные стандарты метаданных не позволяют представить исключение понятий.

Применительно к поисковому предписанию исключение понятий может привести к потере некоторых релевантных документов, таких как статья, посвященная строительству помещений для деятельности профсоюзов. При любом использовании булевого оператора НЕ при поиске необходимо участие человека для оценки и корректировки стратегии поиска.

Следующие рекомендации относятся к сложным соответствиям:

- когда формулировка соответствия включает в себя комбинацию различных символов, для пояснения синтаксиса следует использовать круглые скобки;
- сложные соответствия обычно дают неточные результаты и поэтому полезны в приложениях с участием человека, чтобы отделить нерелевантные результаты от релевантных;
- соответствия, которые включают отрицание или исключение определенных понятий, не рекомендуется использовать, кроме случаев, когда результаты поиска могут быть оценены и стратегия поиска скорректирована.

14 Методы установления соответствий

14.1 Общая процедура

Традиционно идентификация соответствий является интеллектуальным процессом. Требуется один или несколько экспертов, знакомых с соответствующей предметной областью (областями), свободно владеющих языком (языками) словарей, которые необходимо сопоставить, и хорошо понимающих структуру и условные обозначения этих словарей. Процедура заключается в следующем:

a) если сопоставляется более двух словарей, следует определить наиболее подходящую структурную модель (см. раздел 6). После того, как это сделано, работа обычно распределяется так, что каждый эксперт сопоставляет только одну пару словарей;

b) необходимо выбрать направление соответствий, а также диапазон используемых типов соответствий. Если соответствия необходимы в обоих направлениях, может оказаться удобным подготовить все соответствия систематически сначала в одном направлении, а затем определить все соответствия в противоположном направлении;

c) для повышения эффективности и тщательности работы эксперта исходный словарь должен быть доступен в последовательности, которая поддерживает систематическую работу, например, в иерархиях, организованных по дисциплинам или отраслям знания. Последовательность дисциплин может быть организована в зависимости от наличия ресурсов и приоритетности;

d) эксперт должен систематически прорабатывать понятие за понятием целевого словаря, чтобы установить нужное соответствие (соответствия);

e) в большинстве базовых систем эксперту может потребоваться вводить каждое утверждение о соответствии вручную. Если это делается таким образом, каждый термин должен автоматически проверяться на соответствие исходному и целевому тезаурусу, чтобы обеспечить верную орфографию дескрипторов. (Если один из словарей является классификационной системой, проверка применяется к классификационному коду, а не к термину или наименованию класса. В случае словарей предметных рубрик проверяются рубрики. В тех случаях, когда допустимы синтезированные классификационные

коды или новые предкоординированные рубрики, требуется разработка алгоритма автоматической проверки.) В более сложных системах ввод данных может быть полностью или частично автоматизирован, например, путем заполнения шаблона всеми дескрипторами из исходного тезауруса в систематической последовательности и/или путем получения соответствий с помощью выбора на экране терминов целевого тезауруса. При всех формах автоматизации важно дать возможность эксперту отменить неправильный выбор терминов, которые были введены автоматически;

f) если требуется соответствие в противоположном направлении, роли исходного и целевого словаря следует поменять местами и повторить ту же процедуру. Существующие соответствия (из Словаря А в Словарь В) могут быть использованы при соответствиях из В в А следующим образом:

1) когда была надежно идентифицирована точная эквивалентность, то же самое соответствие может быть введено в противоположном направлении автоматически.

Пример — Если термин «антенны» в Словаре А имеют соответствие точной эквивалентности (см. 11.2) термину «антенна» в Словаре В, обратное соответствие «антенна =ЭК антенны» может быть принято без проверки;

2) когда установлена неточная эквивалентность, эксперту для проверки и утверждения или отклонения/модификации может быть предложено такое же соответствие в обратном направлении.

Пример — Если термин «возвращение права владения» в Словаре А имеет неточное (см. 11.3) соответствие эквивалентности термину «право выкупа» в Словаре В, обратное соответствие «право выкупа ~ЭК возвращение права владения» может быть предложено эксперту для проверки и утверждения или отклонения/модификации;

3) когда установлена пересекающаяся сложная эквивалентность (см. 8.3.2), эксперту может быть предложено два или более отдельных иерархических соответствия от более широкого понятия к более узкому в обратном направлении для проверки и утверждения или отклонения/модификации.

Пример — Если «пловцы-призеры» Словаря А были сопоставлены с «призеры + пловцы» в Словаре В, можно предложить эксперту для проверки и утверждения или отклонения/модификации такие обратные соответствия: «призеры УС пловцы-призеры» и «пловцы УС пловцы-призеры»;

4) когда установлена кумулятивная сложная эквивалентность (см. 8.3.3), эксперту может быть предложено два или более отдельных иерархических соответствия от более узкого понятия к более широкому в обратном направлении для проверки и утверждения или отклонения/модификации.

Пример — Если «воздушные суда» Словаря А сопоставлены с «самолеты \| вертолеты» в Словаре В, для проверки и утверждения или отклонения/модификации эксперту могут быть предложены обратные соответствия «самолеты ШС воздушные суда» и «вертолеты ШС воздушные суда»;

5) когда твердо установлено иерархическое соответствие (см. раздел 9) от более узкого понятия к более широкому, может быть предложено эксперту для проверки и утверждения или отклонения/модификации иерархическое соответствие от более широкого понятия к более узкому.

Пример — Если термин «булыжники» Словаря А имеет иерархическое соответствие (от узкого к широкому) термину «камни» в Словаре В, обратное соответствие «камни УС булыжники» может быть предложено эксперту для проверки и утверждения или отклонения/модификации;

6) когда твердо установлено иерархическое соответствие (см. раздел 9) от более широкого понятия к более узкому, иерархическое соответствие от более узкого понятия к более широкому может быть предложено эксперту для проверки и утверждения или отклонения/модификации.

Пример — Если термин «крупный рогатый скот» в Словаре А имеет иерархическое соответствие (от широкого к узкому) термину «коровы» в Словаре В, то для проверки и утверждения или отклонения/модификации эксперту может быть предложено обратное соответствие «коровы ШС крупный рогатый скот»;

7) когда установлено ассоциативное соответствие (см. раздел 10), эксперту для проверки и утверждения или отклонения/модификации может быть предложено такое же соответствие в обратном направлении.

Пример — Если термин «продвижение» в Словаре А имеет ассоциативное соответствие термину «рекламирование» в Словаре В, обратное соответствие «рекламирование АС продвижение» может быть предложено эксперту для проверки и утверждения или отклонения/модификации.

14.2 Автоматизация прямого сопоставления

Иногда процесс, описанный в 14.1 d), можно автоматизировать хотя бы частично.

Если оба словаря являются тезаурусами хотя бы с одним общим языком, процедура состоит в том, чтобы сравнить каждый термин в исходном тезаурусе со всеми терминами на том же языке в целевом словаре. Если совпадения найдены, то они могут быть приняты в качестве потенциальных соответствий. В случаях, когда понятие в исходном словаре выражено непредпочтительными терминами (аскрипторами), для каждого из них могут быть найдены разные потенциальные соответствия, как и для предпочтительного термина (дескриптора).

Дополнительные потенциальные соответствия можно найти, убрав все реляторы терминов и повторив процесс.

Если тезаурусы имеют более одного общего языка, процесс должен повторяться для каждого из них.

Если потенциальное соответствие для понятия еще не было найдено каким-либо из вышеперечисленных способов, из терминов как на целевом, так и на исходном языках могут быть выделены словообразовательные основы, после чего процесс нахождения соответствия следует повторить для каждого языка. Могут применяться дополнительные методы обработки естественного языка, но их описание выходит за рамки этого стандарта.

Потенциальные соответствия, выявленные описанными способами, должны быть собраны для рассмотрения экспертом. Для каждого понятия в исходном словаре эксперт должен иметь возможность просматривать полную запись (включая лексическое примечание, более широкие и более узкие термины). Потенциальные соответствия для понятия должны быть представлены в ранжированном порядке, чтобы эксперту удобно было в первую очередь проводить сравнения и находить точные эквиваленты. Например, совпадения по предпочтительным терминам обычно имеют самый высокий рейтинг, совпадения по непредпочтительным терминам — более низкий, затем следуют совпадения после исключения реляторов и, наконец, совпадения после выделения основ. Интерфейс просмотра должен облегчать проверку полного контекста каждого понятия, определенного в целевом словаре. Он также должен помогать эксперту в выборе соответствующего типа соответствия для варианта (вариантов), который он утверждает.

Только после того как эксперт рассмотрит все потенциальные соответствия, они могут быть утверждены и установлены.

Если один из сопоставляемых словарей не является тезаурусом, процедура соответствия должна быть адаптирована. Наименования классов в классификационной системе или категориальные метки в таксономии гораздо менее надежны, чем термины тезауруса, для однозначного описания класса или категории (см. 17.2.2). В случае словаря предметных рубрик рубрики, состоящие из одного термина, могут рассматриваться как термины тезауруса, но сложные рубрики нуждаются в особой обработке (см. 20.3).

Описанные процедуры соответствия часто могут быть расширены или адаптированы в соответствии с контекстом конкретного словаря или приложения.

14.3 Соответствия на основе совместной встречаемости

Этот метод использует совместную встречаемость терминов из разных систем в метаданных или записях каталога. Его применение предполагает, что существует хотя бы одна большая коллекция, уже проиндексированная с помощью обоих словарей. Принимая Словарь А в качестве исходного словаря, анализируют метаданные всех записей, проиндексированных определенным термином, чтобы определить, какие термины из Словаря В встречаются совместно с ним. Термины, имеющие наивысшую частотность, могут быть использованы в качестве вариантов для обоснованного выбора соответствий. Любое из потенциальных соответствий может быть утверждено и установлено только после рассмотрения экспертом. Описание алгоритма определения совместной встречаемости выходит за рамки данного стандарта.

14.4 Другие методы

В данной быстро развивающейся области могут появиться и другие методы. Настоящий стандарт не исключает использование каких-либо новых технологий, однако рекомендуется для получения надежных соответствий хорошего качества любые потенциальные соответствия, которые генерируются автоматически, передавать на рассмотрение эксперту.

15 Управление данными

15.1 Типы данных

15.1.1 Обзор

Данные сопоставлений могут быть представлены на трех уровнях:

a) данные, описывающие сопоставление отдельных понятий;

b) данные, описывающие сопоставление групп понятий между всеми или частью двух контролируемых словарей (например, полный набор соответствий между двумя тезаурусами или набор соответствий одного микротезауруса словарю предметных рубрик, или соответствия одной группы специфических понятий в тезаурусе другому словарю).

Примечание — В некоторых случаях набор соответствий может включать комбинацию из двух или более целевых словарей, где один или несколько из них предоставляют реляторы для терминов других словарей (см. 8.3.4);

c) данные, описывающие кластер соответствий, то есть скоординированный набор соответствий между более чем двумя контролируемыми словарями (например, набор соответствий между одним центральным словарем и тремя другими словарями, как показано на рисунке 2 в 6.4).

15.1.2 Соответствия между отдельными понятиями

На этом уровне данные должны включать:

a) два (или более) идентификатора понятий, дескриптора или классификационных кода, которые должны быть сопоставлены [всего два в случае простого соответствия и более двух в случае сложной эквивалентности (см. 8.3)];

b) тип соответствия между понятиями (см. разделы 7—10), включающий, если нужно, составные операторы ('|' или '+'), объединяющие целевые понятия (см. 8.3), и маркер точности или неточности (см. раздел 11);

c) период действия соответствия, если он не подразумевается в спецификации сопоставляемых словарей.

Примечание — Это требуется, когда термины в контролируемом словаре добавляются, удаляются или изменяются в объеме понятия и когда нужно учитывать какие-либо исторические справки к отдельным понятиям в соответствующих словарях;

d) возможные пояснения с дополнительной необязательной информацией (например, о рейтинге или о степени уверенности в качестве соответствия).

Примечание — Полный перечень таких пояснений зависит от конкретного приложения и выходит за рамки данного стандарта.

Сопоставление отдельных понятий имеет смысл только в контексте контролируемых словарей, в которых объем этих понятий определен явно или неявно, поэтому эти словари должны быть идентифицированы, как указано в 15.1.3.

Для каждого соответствия также должны быть определены характеристики, описанные в 15.1.3 b)—d). Однако может быть удобнее указывать их на уровне группы, а не индивидуально. Для некоторых соответствий может быть целесообразным также указать на индивидуальной основе характеристики, описанные в 15.1.3 e)—j).

15.1.3 Набор соответствий между группами понятий

На уровне соответствий между группами понятий, охватывающими целиком или частично два контролируемых словаря, данные должны указывать:

a) идентификацию набора соответствий, включая, где это уместно, указание версии или даты последнего изменения соответствий;

b) исходный словарь (включая его дату и/или версию);

c) целевой словарь (включая его дату и/или версию);

d) направление соответствий, если они односторонние;

e) любые ограничения на область сопоставления; например, если исходный и целевой словари не включены в группу целиком, то должны быть указаны используемые части или принцип выбора; если соответствия были разработаны для конкретной цели или применения, это должно быть указано;

f) процесс, с помощью которого были созданы соответствия. Например, с помощью анализа человеком, автоматического сопоставления терминов, на основе совместной встречаемости в коллекции ресурсов, проиндексированных с помощью обоих словарей, или некоторой комбинации таких методов;

g) идентификацию лица (лиц) или организации, ответственной за создание и/или публикацию соответствий;

h) как можно получить доступ к соответствиям (см. также 15.3.4);

i) связанные соответствия или кластеры соответствий, к которым принадлежит этот набор соответствий, как описано в 15.1.4;

j) ограничения авторского права, конфиденциальности или безопасности на использование или обмен соответствиями.

По желанию могут быть добавлены примечания, касающиеся степени уверенности в качестве набора соответствий.

15.1.4 Кластеры соответствий

Если сопоставление формирует часть кластера соответствий между более чем двумя контролируемыми словарями, должны быть указаны следующие данные о кластере:

a) идентификация кластера;

b) структурная модель кластера (см. раздел 6) и, в случае централизованной структуры, идентификация центрального контролируемого словаря;

c) предполагаемое назначение или применение кластера;

d) идентификация лица (лиц) или организации, ответственных за поддержание и/или публикацию кластера; а также

e) ограничения на использование или обмен кластером, связанные с авторским правом, конфиденциальностью или безопасностью.

По желанию могут быть добавлены примечания, касающиеся степени уверенности в качестве кластера соответствий.

15.2 Хранение данных

15.2.1 Организационные аспекты

Соответствие может быть сохранено:

- в системе, управляющей исходным контролируемым словарем;
- в системе, управляющей целевым контролируемым словарем;
- в системе, предназначенной для управления соответствиями, но не зависящей от какой-либо системы, управляющей целевым или исходным контролируемыми словарями; или же
- в системе, управляющей всеми исходными и целевыми контролируемыми словарями, а также соответствиями между ними.

Обычно выбор организационного режима определяется владельцем данных и бизнес-моделью, а не техническими критериями. Тем не менее некоторые последствия этого выбора рассматриваются в 15.3.

15.2.2 Технические аспекты

Соответствия, наборы соответствий и кластеры обычно хранятся в базе данных. Для каждого соответствия должны быть определены характеристики, указанные в 15.1. Каждое отдельное соответствие должно быть детализировано как отношение или как правило.

Хранение этих соответствий может быть выполнено с использованием технологий, таких как реляционная база данных, база данных XML, база данных правил или хранилище RDF. Каждый из этих методов требует применения отдельной схемы, удовлетворяющей следующим требованиям:

- схема должна иметь возможность указывать различные типы соответствия, описанные в разделах 7—10, включая сложную эквивалентность (см. 8.3), если она используется;
- схема должна иметь возможность указывать примечания к соответствиям;
- схема должна иметь возможность задавать характеристики наборов соответствий (см. 15.1.3);
- схема должна иметь возможность указывать характеристики кластеров соответствий (см. 15.1.4).

Пока еще нет стандартной схемы, которая бы полностью соответствовала этим требованиям, и разработка такой структуры в настоящем стандарте пока не предлагается. Однако это направление быстро развивается, и при внедрении стандарта нужно быть готовым к появлению заинтересованности в новых разработках, например, среди пользователей SKOS [37].

Схемы, предназначенные для хранения, также могут использоваться или адаптироваться для обеспечения возможности публикации соответствий. Для использования в Семантическом вебе рекомендуется SKOS-совместимый формат [37].

15.3 Сохранение данных о соответствиях

15.3.1 Общие положения

Поддержание актуальности данных о соответствиях является сложной задачей. Следует проявлять осторожность при использовании инструментов и процедур, которые могут обеспечить это. Во всех случаях инструменты должны иметь доступ к исходному и целевому словарям.

При ведении контролируемого словаря, если словарь связан с набором или кластером соответствий, нельзя допускать изменения объема понятий, связанных с соответствиями, без внесения также изменений в эти соответствия. *Типы возможных изменений указаны в ГОСТ Р 7.0.91—2015, пункт 13.6.4.*

15.3.2 Изменения в исходных или целевых словарях

Когда исходный или целевой словарь изменяется, должны быть проверены прямые и обратные соответствия для его понятий. *Список различий между предыдущей версией и новой версией словаря, как это рекомендовано в ГОСТ Р 7.0.91—2015, подпункты 13.6.5.2 и 13.6.5.3, может помочь процессу проверки, особенно если определены изменения, которые влияют на объем сопоставленных понятий.* Автоматические тесты могут быть ненадежными, поскольку они обычно не указывают, например, является ли новый непредпочтительный термин (аскриптор) просто вариантом написания или представляет собой расширение объема соответствующего понятия.

После того как все соответствия были скорректированы, должна быть выпущена новая версия набора или группы соответствий.

В таблице 7 показаны основные типы действий, требуемых во время или после обновления исходного словаря, а таблица 8 применяется во время или после обновления целевого словаря.

Таблица 7 — Последствия изменений в исходном словаре

Изменения в исходном словаре	Корректировка соответствий
Введение нового понятия	Требуется установить новое соответствие, если в целевом словаре есть соответствующее понятие
Удаление понятия	В исходном словаре соответствие данному понятию следует либо снабдить пояснением о периоде действия, чтобы указать ограничение на его использование; либо удалить из новой версии набора или кластера соответствий
Комплексная замена понятия (например, разделение или слияние)	Все релевантные соответствия должны быть проверены и исправлены, если это необходимо. Они не должны обновляться автоматически
Изменение объема понятия	Существующие соответствия от этого понятия должны быть проверены и исправлены, если это необходимо

Таблица 8 — Последствия изменений в целевом словаре

Изменения в целевом словаре	Корректировка соответствий
Введение нового понятия	Существующие соответствия остаются в силе, но, вероятно, нуждаются в уточнении. Должно быть проверено наличие иерархических и ассоциативных соответствий для нового понятия. В исходном словаре следует найти все понятия, которые можно было бы с пользой отобразить на новое целевое понятие
Удаление понятия	Все простые соответствия удаленному целевому понятию должны быть удалены и, при необходимости, заменены наиболее подходящими соответствиями оставшимся исходным понятиям Все сложные соответствия с удаленным целевым понятием должны быть изучены и исправлены или заменены, если это необходимо
Комплексная замена понятия (например, разделение или слияние)	Все соответствия, затронутые изменениями, должны быть проверены и исправлены или заменены новыми, если это необходимо. Они не должны обновляться автоматически
Изменение объема понятия	Существующие соответствия к этому понятию должны быть проверены и исправлены, если это необходимо

15.3.3 Другие изменения в соответствиях

Необходимость пересмотреть и скорректировать соответствия может возникнуть, даже если исходный и целевой словари остаются неизменными. Если соответствия каким-либо образом пересматриваются, об этом следует сообщать пользователям и отражать изменения во всех информационно-поисковых системах, которые зависят от этих соответствий. Следует помнить, что отдельное соответствие может принадлежать более чем одному набору или группе соответствий, и все наборы, затронутые изменениями, должны быть обновлены соответствующим образом.

15.3.4 Влияние изменений в соответствиях

Поскольку соответствия и наборы соответствий разрабатываются главным образом для поиска информации, как описано в разделе 5, любые изменения в них должны отражаться в программах индексирования и поиска, которые их используют (см. также пояснения и примеры в разделе 12). Должен быть составлен перечень изменений, затрагивающих каждую группу или набор соответствий, который должен обеспечивать:

a) реконфигурацию всех механизмов индексирования, которые используют соответствия;

b) оценку влияния на метаданные документов, уже проиндексированных с помощью соответствий;

c) реконфигурацию всех поисковых систем, использующих эти соответствия. В тех случаях, когда соответствия применяются только к поисковым терминам (а не к индексным терминам), негативное влияние на эффективность поиска маловероятно.

Если соответствия опубликованы, список изменений также должен быть опубликован.

Как правило, проходит значительное время, прежде чем весь проиндексированный контент будет актуализирован новейшими версиями обновленных словарей и наборов соответствий. Поэтому более старые версии словарей и наборов соответствий должны оставаться доступными до тех пор, пока они больше не будут использоваться (иногда соответствия нуждаются в пояснениях, которые ограничивают их во времени или предназначены для конкретных версий; см. 15.1.2 d).

16 Визуализация сопоставленных словарей

16.1 Общие положения

Ни один способ представления сопоставленных словарей не отвечает всем вероятным потребностям. Для поисковых приложений постоянно появляются новые стили визуализации. Этот стандарт не ограничивает способы презентации данных о сопоставлении словарей конечным пользователям, а дает только общие принципы. Визуализации, описанные в 16.2—16.4, предназначены главным образом для управления установлением и ведением системы соответствий.

Для некоторых целей достаточно представить только один из отображенных словарей (без раскрытия соответствий), и, если это тезаурус, представление должно быть таким, как описано в ГОСТ Р 7.0.91—2015, раздел 12.

От конечных пользователей не следует ожидать, что они смогут распознать и правильно понять метки, описанные в разделе 4. Метки предназначены для использования обученным персоналом.

В некоторых приложениях сопоставления словарей используются в основном компьютером, и их визуализация может оказаться затруднительной и ненужной. Например, людям вообще не нужно видеть соответствия, если термины, которые они вводят в поисковую систему, автоматически конвертируются в эквиваленты на другом языке. Даже если предусмотрено участие человека (см. 12.2), варианты могут быть представлены пользователю таким способом, который не требует от него знания процесса сопоставления. Например, сопоставленные термины из целевого словаря могут быть представлены в виде облака тэгов без явного обозначения типа соответствия в каждом случае.

Тем не менее, визуализация соответствий может быть полезна для просмотра доступных словарей. И какая-то форма визуализации может быть предоставлена пользователям, когда нужно выбирать между альтернативными соответствиями.

В приложениях, где поощряется выбор пользователем поисковых элементов (терминов или классификационных кодов), должна предоставляться поддержка прямого доступа к терминам словаря, навигация по гиперссылкам, удобные для пользователя иерархические представления и другие визуальные подсказки, насколько позволяют ресурсы.

Когда для выбора представлено большое количество соответствий, они должны быть сгруппированы в удобные группы, чтобы избежать путаницы. В зависимости от приложения и контекста группи-

ровка может осуществляться по типу соответствия, по степени эквивалентности и/или степени перекрытия (в случае неточной эквивалентности), по целевому словарю, по типу целевого объекта (в случае нормативных файлов имен) или по какому-либо другому атрибуту, который важен для пользователей.

Когда исходные или целевые словари представлены более чем на одном языке, интерфейс просмотра должен позволять пользователю переключать язык по своему выбору, и желательно, чтобы все соответствующие термины были доступны на выбранном языке.

Визуализация соответствий также должна быть доступна лицам, которые занимаются их установлением и поддержкой. Представления, показанные в 16.2 и 16.3, могут оказаться полезными в этом контексте.

16.2 Представление одной словарной статьи

Для пользователей, обученных использованию соответствий, часто требуется представление одной словарной статьи, особенно в процессе создания и ведения соответствий. В этом типе представления один словарь принимается в качестве исходного, из которого установлены или устанавливаются соответствия с целевым словарем (словарями).

Представление должно соответствовать положениям ГОСТ Р 7.0.91—2015, подраздел 12.2, с дополнительным указанием на соответствия терминам целевого словаря с использованием условных обозначений, установленных в разделе 4 настоящего стандарта и более подробно разъясненных в разделах 7—13. Необязательные элементы исходного словаря (например, наивысший термин и определение) могут быть включены при необходимости.

Стиль представления зависит от того, насколько различаются виды соответствия и степени эквивалентности при сопоставлении. Варианты представления, основанные на данных, извлеченных из трех тезаурусов в таблице 9, показаны в таблицах 10 и 11. Соответствия здесь устанавливаются от одного исходного тезауруса к двум целевым. В первом целевом тезаурусе есть понятие с полной эквивалентностью, но в другом имеются только варианты либо намного более широкие, либо намного более узкие, которые, возможно, подходят для установления иерархических соответствий.

Таблица 9 — Сопоставленные статьи трех тезаурусов, для которых строятся соответствия

Исходный словарь	Целевой словарь 1	Целевой словарь 2
молочные продукты	молокопродукты	животные продукты
с *молокопродукты*	в животные продукты	в продукты
в животные продукты	продукты питания	н кожа
н масло	н мороженое	молоко
молоко	сыры	мясо
сливки	а молокозаводы	а животноводство
сыр		
а молокозаводы		

В этой ситуации сопоставление может быть выполнено с большей или меньшей дифференциацией, как показано в таблице 10. Словарь 1 и Словарь 2 — это два целевых тезауруса. В первой колонке (недифференцированное соответствие) показан минимальный подход, при котором для понятия из исходного словаря было определено только одно соответствие в целевом словаре, а тип соответствия не указан. Во второй колонке (дифференцированное соответствие) были предприняты дополнительные усилия для идентификации и представления всех соответствий, которые могут быть полезны для конкретного понятия. Последний подход поддерживает различие между альтернативными соответствиями, когда точное эквивалентное понятие отсутствует. Этот стиль сопоставления должен также учитывать имеющиеся сложные соответствия эквивалентности.

Таблица 10 — Представление одной записи для двух разных стилей соответствия (от одного исходного тезауруса к двум целевым тезаурусам в обоих случаях)

Недифференцированное соответствие	Дифференцированное соответствие
молочные продукты	молочные продукты
с *молокопродукты*	с *молокопродукты*
в животные продукты	в животные продукты
н масло	н масло
молоко	молоко
сливки	сливки
сыр	сыр
а молокозаводы	а молокозаводы
Словарь 1 молокопродукты	Словарь 1 =ЭК молокопродукты
Словарь 2 животные продукты	Словарь 2 ШС животные продукты
	Словарь 2 УС молоко

В таблице 11 возможности представления еще более расширены. К понятиям Словаря 1 добавлены дополнительные действующие соответствия. Иерархические и/или ассоциативные соответствия понятиям «животные продукты», «пищевые продукты», «сыры», «мороженое» и «молочные продукты» не ошибочны, но их обычно считают ненужными устанавливать и указывать, поскольку они могут быть автоматически выведены из структуры внутренних отношений в рамках Словаря 1, при условии установления точной эквивалентности понятию «молокопродукты».

Таблица 11 — Представление одной записи с дифференцированным соответствием, расширенное для включения избыточных соответствий

молочные продукты		
с	*молокопродукты*	
в	животные продукты	
н	масло	
	молоко	
	сливки	
	сыр	
а	молокозаводы	
Словарь 1	=ЭК	молокопродукты
Словарь 1	ШС	животные продукты
		пищевые продукты
Словарь 1	УС	мороженое
		сыры
Словарь 1	АС	молокозаводы
Словарь 2	ШС	животные продукты
Словарь 2	УС	молоко

Дифференцированный стиль без избыточности, показанный во второй колонке таблицы 10, рекомендуется в качестве наиболее мощного и гибкого варианта.

Недифференцированный стиль может быть приемлемым в качестве недорогой альтернативы в приложениях, которые не требуют соответствий высокого качества.

Как показано в таблицах 10 и 11, метка для обозначения вида соответствия в одноязычном контексте имеет три компонента: во-первых, идентификатор целевого тезауруса, во-вторых — обозначение типа соответствия и, факультативно, — знак тильды или равенства, чтобы показать неточную или точную эквивалентность, где это применимо. *Все эти метки и символы объяснены в разделе 4 настоящего стандарта, а также в ГОСТ Р 7.0.91—2015, раздел 3*. Идентификаторы исходного и целевого тезаурусов следует выбирать так, чтобы они не путались с другими обозначениями.

Если целевой тезаурус является многоязычным, его идентификатор может дополнительно включать указатель языковой версии. В этом случае ***метка «Словарь 1» должна стать «Словарь 1 рус», а «Словарь 1 ЭК» — стать «Словарь 1 рус ЭК».***

Если рассматриваются соответствия с другим типом словарей (вместо или вместе с ***«Словарь 1» и «Словарь 2»*** в таблицах 9—11), в каждом соответствии должны быть показаны следующие целевые компоненты:

- для словаря предметных рубрик — соответствующая рубрика (рубрики);
- для нормативного файла имен — подходящее предпочтительное имя (имена) или идентификатор (идентификаторы) объекта;
- для классификационной системы — соответствующие коды и наименования классов;
- для таксономии — соответствующая метка (метки) категории и, если метка не является уникальной в таксономии, некоторые средства устранения неоднозначности.

Если соответствия предназначены для использования в разных направлениях, для каждого словаря, который будет применяться в качестве исходного, должно быть доступно аналогичное представление.

16.3 Полное представление на базе одного из словарей

16.3.1 Алфавитное представление

Если исходный словарь, выбранный для алфавитного представления, является тезаурусом, отдельные записи в полном отображении должны быть такими, как описано в 16.2. Если все остальные словари являются одноязычными тезаурусами, то полное представление будет таким, как показано на рисунке 9. Этот рисунок основан на тех же трех тезаурусах, которые использованы в таблице 9. Для каждого тезауруса (т. е. для Словаря 1 и Словаря 2 в этом примере), когда он используется в качестве исходного словаря, должно быть доступно аналогичное представление.

Если исходный словарь является нормативным файлом имен, представление по алфавиту может быть подготовлено аналогичным образом.

агрохимикаты
с *сельскохозяйственные химикаты*
н удобрения
пестициды
Словарь 1 =ЭК агрохимические продукты
Словарь 2 =ЭК сельскохозяйственные химикаты

животные продукты
н молочные продукты
яйца
кожа
мясо
шерсть
Словарь 1 =ЭК животные продукты
Словарь 2 =ЭК животные продукты

злаки
в растительные продукты
Словарь 1 =ЭК злаки
Словарь 2 ШС растительные продукты

крупный рогатый скот
в домашний скот
а молоко
Словарь 1 =ЭК крупный рогатый скот
Словарь 2 ~ЭК коровы

масло
в молочные продукты
Словарь 1 ШС молокопродукты
Словарь 2 ШС животные продукты
Словарь 2 АС молоко

молочные продукты
в животные продукты
н масло
сыр
сливки
молоко
а молокозаводы
Словарь 1 =ЭК молокопродукты
Словарь 2 ШС животные продукты
Словарь 2 УС молоко

сливки
в молочные продукты
а молоко
Словарь1 ШС молокопродукты
Словарь 2 ШС животные продукты
Словарь 2 АС молоко

сыр
в молочные продукты
Словарь 1 =ЭК сыры
Словарь 2 ШС животные продукты
Словарь 2 АС молоко

цыплята
с куры
в домашняя птица
Словарь 1 =ЭК цыплята
Словарь 2 =ЭК цыплята

Рисунок 9 — Алфавитное представление с соответствиями в двух целевых словарях

16.3.2 Систематическое представление

Систематическое представление соответствий может быть полезно, когда исходный тезаурус использует преимущественно классификационное представление. Оно также применимо, когда исходный словарь представляет собой классификационную структуру (включая системы, используемые для управления документами), таксономию или словарь предметных рубрик, поскольку они обычно представлены в систематическом, а не в алфавитном порядке.

16.4 Сопоставительные таблицы

Для многих приложений визуализация внутренних связей исходного или целевого словарей не требуется. Сопоставительная таблица, показывающая только соответствия между словарями, может оказаться более удобной. В таблице 12 показаны соответствия одного исходного тезауруса двум целевым тезаурусам на основе тех же данных, что и в таблице 9 и на рисунке 9.

Таблица 12 — Таблица соответствий в двух целевых словарях

Понятие исходного тезауруса	Соответствие в Словаре 1		Соответствие в Словаре 2	
	Тип соответствия	Понятие	Тип соответствия	Понятие
агрохимикаты	=ЭК	агрохимические продукты	=ЭК	сельскохозяйственные химикаты
животные продукты	=ЭК	животные продукты	=ЭК	животные продукты
злаки	=ЭК	злаки	ШС	растительные продукты
крупный рогатый скот	=ЭК	крупный рогатый скот	~ЭК	коровы
масло	ШС	молокопродукты	ШС АС	животные продукты молоко
молочные продукты	=ЭК	молокопродукты	ШС УС	животные продукты молоко
сливки	ШС	молокопродукты	ШС ШС	животные продукты молоко
сыр	=ЭК	сыры	ШС АС	животные продукты молоко
цыплята	=ЭК	цыплята	=ЭК	цыплята

При интерпретации данных в такой таблице есть важное ограничение, о котором следует сказать. В каждом случае соответствия устанавливаются между понятием в колонке 1 (взятым из исходного тезауруса) и соответствующими понятиями в колонках 3 и 5. Если соответствия не относятся к типу точной эквивалентности, то эта таблица не показывает соответствий одного целевого словаря другому целевому словарю. Например, в третьем с конца таблицы ряду мы видим понятие Словаря 1 «молокопродукты» в той же строке, что и понятие Словаря 2 «животные продукты» и «молоко», но без дополнительной информации тип соответствия этих понятий не может быть точно определен.

17 Классификационные системы

17.1 Ключевые характеристики и происхождение

17.1.1 Общее описание

Классификационная система — это система классов или категорий, предназначенная для систематизации информационных ресурсов любых видов. Как описано в этом разделе, ее можно использовать для систематизации единиц хранения как на библиотечных полках, так и в каталогах. Классификационные структуры, используемые для управления документами, и для организации веб-страниц и порталов, описаны отдельно в разделах 18 и 19 соответственно.

Основной метод классификации состоит в том, чтобы организовать понятия в классы. Классы подразделяются на более узкие классы (см. 17.2.3), которые делятся на еще более узкие классы вплоть до требуемого уровня детальности.

Каждый класс имеет наименование (см. 17.2.2), а иногда и более длинное описание, которое определяет содержание класса. Функционируя как лексическое примечание тезауруса, описание класса часто перечисляет входящие в класс темы и включает перекрестные ссылки «см. также» на другие классы, в которых находятся связанные темы.

В тематических классификационных системах (см. 17.1.5.1) каждый класс имеет также обозначение (см. 17.2.1), обычно короткую строку буквенных и/или цифровых и других символов.

Классификационный подход обычно предусматривает предкоординацию понятий. Он принят во многих других типах словарей, включая многие таксономии (см. раздел 19) и системы управления документами (см. раздел 18). Поскольку предкоординация создает проблемы для совместимости, последствия и примеры для всех таких словарей рассматриваются далее в разделе 19.

Некоторые примеры различных стилей классификационных систем показаны на рисунках 10 и 11. В обоих случаях показаны только краткие фрагменты, содержащие наименования (см. 17.2.2) и коды классов (см. 17.2.1), а прочая информация исключена.

17.1.2 Место и роль в информационном поиске

Цель классификационных систем традиционно состоит в том, чтобы организовать информационные ресурсы, будь то в печатном или электронном виде, и обеспечить поиск путем просмотра полок или систематического каталога.

Современные информационные технологии позволяют расширить использование классификационных систем для поиска по множеству коллекций и баз данных.

17.1.3 Создание и развитие

Классификационные системы использовались на протяжении всей истории библиотек. Многие из используемых в настоящее время систем основаны на определенной структуре и методах, разработанных сто и более лет назад, например, ***Десятичная классификация Дьюи*** (ДКД) [16], ***Расширенная классификация*** Каттера [15], ***Универсальная десятичная классификация*** (УДК) Лафонтена и Отле, ***Классификация Библиотеки Конгресса*** (КБК) [21] Хансона и Мартела, ***Предметная классификация*** Брауна [14]. Эти ранние системы были в основном перечислительными (см. 17.1.5.2), хотя УДК ввела некоторые синтетические принципы с самого начала. Идея фасетной классификации появилась позже, с ***Классификацией двоеточием*** Ранганатана (КДР) [27] и ***Библиографической классификацией*** Блисса (БКБ) [11]. С 1950-х годов некоторые из методов фасетной классификации постепенно внедряются в большинство перечислительных систем, в частности, в *Библиотечно-библиографическую классификацию, принятую в большинстве библиотек России.*

17.1.4 Словарный контроль

В то время как тезаурус осуществляет словарный контроль, назначая предпочтительный термин для каждого понятия, классификационная система дает уникальное обозначение для каждого класса. Будучи независимым от языка (языков), классификационный код служит однозначной меткой для представления содержания класса.

Наличие алфавитного указателя (см. 17.2.4) обычно обеспечивает дополнительные точки входа в классы из альтернативных терминов и способов выражения соответствующих понятий. Указатель обеспечивает доступ на естественном языке к классификационным кодам, которые составляют искусственно контролируемый язык системы.

17.1.5 Типы классификационных систем

17.1.5.1 Общие положения

Тип классификационных систем, используемых для управления документами, рассматривается в разделе 18, а тип, используемый для организации веб-страниц и порталов, — в разделе 19. В настоящем разделе рассматриваются главным образом классификационные системы, используемые библиотеками (физическими или электронными) или библиографическими базами данных, где целью системы является предоставление доступа к документам в соответствии с их тематикой. Такие системы иногда называют «тематическими классификационными системами».

17.1.5.2 Перечислительные и синтетические системы

Все тематические классификационные системы представляют собой некоторое количество классов в таблице или наборе таблиц. Процесс установления классов часто называют перечислением. Система, которая полностью предъявляет все доступные пользователю классы, называется перечислительной системой.

Когда в системе нужно указать класс, объединяющий два или более понятия, принадлежащих к разным классам, необходимо принять решение о том, где его разместить. Обычно он находится под одним из составляющих классов в соответствии с правилами, специфичными для рассматриваемой системы.

В качестве альтернативы подробному перечислению, некоторые системы включают в себя правила объединения классов, так что пользователь может синтезировать составные классификационные коды из элементарных кодов составляющих классов. Таким образом, система может предусмотреть гораздо больше тем, чем можно было бы перечислить. Системы с такой возможностью известны как синтетические или аналитико-синтетические системы.

Если классификационные коды рассматривать как составляющие основного контролируемого словаря классификационной системы, то в перечислительной системе этот словарь конечен и изложен полностью. В синтетической системе словарь определен менее строго. Форму и структуру синтезиро-

ванных кодов определяют правила синтеза. Обычно синтезированные коды можно разбить на составляющие, перечисленные в таблицах (см. также 13.2.1)

Часто классификационные системы сочетают перечислительные и синтетические признаки.

17.1.5.3 Фасетные классификационные системы

Фасетные классификационные системы продвигают принципы анализа и синтеза еще на один шаг вперед. Сначала темы разбивают на простые понятия в соответствии с базовыми категориями, такими как «вид деятельности», «предмет», «место» и т. п. Эти понятия перечисляются в системе как классы, каждому из которых присваивается код. Затем коды комплексных тем синтезируют путем соединения кодов простых понятий в соответствии с правилами очередности, известными как фасетная формула. Правила очередности необходимы для того, чтобы все документы по одной и той же комплексной теме имели одинаковые коды и, следовательно, объединялись. Рисунок 10 иллюстрирует простую систему фасетной классификации, в которой используются два фасета: «организмы» и «процессы».

Классификационная таблица	
Код	Наименование
(организмы)	
A	млекопитающие в целом
AA	плотоядные в целом
AAA	леопарды
AAB	львы
AAC	тигры
AB	травоядные в целом
ABA	крупный рогатый скот
ABB	овцы
(процессы)	
B	физиологические процессы в целом
BB	пищеварение в целом
(пищеварение по организмам)	
BB.AA	пищеварение у плотоядных
BB.AAB	пищеварение у львов
BB.AB	пищеварение у травоядных
BB.ABA	пищеварение у крупного рогатого скота
BB.ABB	пищеварение у овец
BC	дыхание в целом
(дыхание по организмам)	
BC.AA	дыхание у плотоядных
BC.AAB	дыхание у львов

Рисунок 10 — Пример короткой фасетной классификации с простыми кодами

17.2 Семантические компоненты и отношения в сопоставлении с компонентами тезауруса

17.2.1 Классификационный код

17.2.1.1 Цели

Основные функции кодов:

a) обеспечить систематическое расположение, позволяющее найти понятия внутри классификационной системы и документы, классифицированные по ее схеме;

b) служить однозначной меткой для предполагаемого понятия.

17.2.1.2 Общее описание

Классификационные коды обычно представляют собой короткие строки буквенных, цифровых или других символов или их сочетания, которые могут использоваться для представления классов. В перечислительных системах показаны все строки обозначений, доступные пользователям. Синтетические и фасетные системы предоставляют правила, которые позволяют пользователям синтезировать

строки кодов для комбинаций понятий, которые не были перечислены в таблицах системы. Примеры кодов показаны на рисунках 10 и 11. На рисунке 11 первые девять классов перечислены в системе; последние два были синтезированы по правилам УДК.

Разделительные знаки, такие как двоеточия, запятые, кавычки и т. д., часто используются в кодах, хотя на практике возможны разные варианты. Для интерпретации любых таких символов необходимо ознакомиться с правилами рассматриваемой системы.

(084)	графические документы
(084.12)	фотографии
5	математика и естествознание
59	зоология
591.1	физиология животных
591.132	пищеварение
599	млекопитающие
599.74	плотоядные
599.742.71	«большие кошки»: львы, тигры
599.742.71:591.132	пищеварение у львов
599.742.71(084.12)	фотографии львов

Рисунок 11 — Выбранные классы из классификационной системы (УДК) с кодами, в которых используются цифры и символы

17.2.1.3 Выводы о сопоставлении классов и тезаурусных понятий

Классы, такие как 591.132 (пищеварение) или 599 (млекопитающие) на рисунке 11, относительно просты, и поэтому для них легко найти эквивалентные понятия, точные или неточные, в тезаурусе с подходящей областью действия (см. разделы 8 и 11). В зависимости от относительной специфики классификации и тезауруса, некоторым простым классам альтернативно может потребоваться иерархическое или ассоциативное соответствие (см. разделы 9 и 10). Однако такой класс, как 599.742.71(084.12), настолько сложен, что вряд ли соответствует единому понятию тезауруса. Такие классы очень распространены в больших системах классификации всех типов, и для их соответствия необходимо использовать сложную эквивалентность (см. 8.3).

17.2.2 Наименования классов

Рисунки 10 и 11 показывают наименование (иногда называемое заголовком класса) и код каждого класса. Наименование кратко передает содержание класса, но оно не обязательно должно быть уникальным, поскольку ожидается, что пользователь будет просматривать его в контексте вышестоящего класса. Например, классы «грибы» и «фрукты» могут быть разделены на подклассы с заголовками «ядовитые» и «съедобные». Ожидается, что пользователь поймет, что понятие «ядовитый» — это действительно ядовитые грибы в первом случае и ядовитые фрукты в другом. Предполагается, что не будет путаницы в расположении классов по этим темам, потому что им даны разные коды. В этом отношении классификационные системы отличаются от тезаурусов, в которых понятия однозначно идентифицируются терминами, а не кодами (*см. ГОСТ 7.0.91—2015, пункт 6.2.1*). При установлении соответствий между классификационной системой и тезаурусом нужно учитывать следующее:

- наименование само по себе, как правило, является недостаточным показателем содержания класса;
- утверждения о соответствии должны быть выражены с использованием кода для представления класса, а не с использованием его наименования.

17.2.3 Иерархии в классификационных системах

Процедура деления классов на более узкие классы, а их на еще более узкие классы, по сути, является иерархической. Иерархический подход часто подчеркивают при визуализации классификационной системы тем, что размер и стиль шрифтов меняют в зависимости от уровня подраздела и используют отступы (как на рисунке 10), чтобы показать, какой достигнут уровень. Однако очень важно отметить, что иерархиями в классификационных системах обычно не управляют так же свободно, как иерархическими отношениями в тезаурусе.

Тезаурус признает отношения иерархическими, только если они являются родо-видовыми, партитивными или инстанциональными (***см. ГОСТ 7.0.91—2015, подраздел 10.2***). Например, поскольку

химический элемент не является ни типом, ни частью, ни элементом химии, отношение «в/н» между этими понятиями неприменимо. Тем не менее, в библиотеке удобно найти книги о химических элементах в разделе «Химия». Поэтому подкласс «химические элементы» может быть помещен на соответствующем уровне в иерархии класса «химия» в классификационной системе для использования в библиотеках. Коды аналогично показывают химические элементы как принадлежащие к химии.

Из этого следует, что когда классы классификационной системы сопоставляют с понятиями тезауруса (даже если каждый класс имеет соответствие точной эквивалентности), обычно иерархические отношения между классами не совпадают с иерархическими отношениями между понятиями тезауруса. Это не должно рассматриваться как проблема. В целях поиска информации для перехода между тезаурусом и классификационной системой необходимы соответствия между понятиями и классами, а не между соответствующими внутренними связями в словарях.

Правила фасетной формулы в классификационной системе предназначены для того, чтобы для каждого сложного класса найти уникальное местоположение в системе. Это удобно в традиционных библиотеках, где для каждого документа должно быть одно (и только одно) место, хотя для документов, относящихся более чем к одному предмету, могут быть записи более чем в одном классе систематического каталога. *См. ГОСТ Р 7.0.91—2015, пункт 10.2.5 для рассмотрения способов, которыми тезаурусы могут обрабатывать альтернативные иерархические местоположения понятий.*

17.2.4 Алфавитные указатели к классификационным системам

Большинство классификационных систем имеют указатель, обеспечивающий доступ к нескольким местоположениям, в которых может встретиться одно понятие внутри системы. Это особенно необходимо для фасетных классификационных систем, потому что, как показано на рисунках 10 и 12, простые понятия, такие как «львы», в виде аспектов разбросаны по сложным темам, где они не являются элементами первостепенного значения.

Алфавитный указатель, соответствующий системе на рисунке 10	
дыхание : физиологические процессы	BC
крупный рогатый скот : пищеварение : физиологические процессы	BB.ABA
крупный рогатый скот : травоядные животные : млекопитающие	ABA
леопарды : плотоядные животные : млекопитающие	AAA
львы : дыхание : физиологические процессы	BC.AAB
львы : пищеварение : физиологические процессы	BB.AAB
львы : плотоядные животные : млекопитающие	AAB
Млекопитающие	A
овцы : пищеварение : физиологические процессы	BB.ABB
овцы : травоядные животные : млекопитающие	ABB
пищеварение : физиологические процессы	BB
плотоядные животные : дыхание : физиологические процессы	BC.AA
плотоядные животные : млекопитающие	AA
плотоядные животные : пищеварение : физиологические процессы	BB.AA
тигры : плотоядные животные : млекопитающие	AAC
травоядные животные : млекопитающие	AB
травоядные животные : пищеварение : физиологические процессы	BB.AB

Рисунок 12 — Выборка записей из цепного указателя

Независимо от того, является ли указатель цепным, как показано на рисунке 12, или использует стиль, рекомендованный [2], его функция состоит только в том, чтобы предоставлять доступ к таблицам системы. Иногда это делается путем включения записей для некоторых аспектов содержания класса, которые могут быть явно не упомянуты в его наименовании или сопроводительных примечаниях. Например, в Десятичной классификации Дьюи записи указателя, ведущие к классу 387.2 (Суда), включают «Транспортные средства на воздушной подушке — океан», «Лодки», «Торговые суда», «Суда (морские)» и т. п. При установлении соответствий между классификационной системой и тезаурусом, такие записи указателя могут оказаться полезными при разъяснении области применения соответствующего класса. Они также могут быть полезны при предложении дополнительных иерархических соответствий

понятиям тезауруса. Однако в любом утверждении о соответствии класс должен быть представлен своим кодом, а не какой-либо из записей указателя.

17.3 Рекомендации по сопоставлению тезауруса и классификационной системы

Содержание класса следует определять путем изучения его наименования (наименований), вышестоящих и подчиненных классов, всех описательных примечаний и соответствующих записей указателя.

Объем каждого понятия тезауруса должен быть установлен с такой же тщательностью (см. также раздел 14).

Соответствия могут быть установлены от классов к понятиям тезауруса или от понятий тезауруса к классам, или в обе стороны. В процессе работы важно принимать во внимание направление соответствия (более подробное рассмотрение и примеры см. в 13.2).

Для простых классов целесообразно найти отдельное понятие тезауруса как точную или неточную эквивалентность (см. 8.2 и раздел 11). Если это не удается, может оказаться возможным установить иерархическое или ассоциативное соответствие (см. разделы 9 и 10). Но для классов, включающих предкоординацию более простых классов, обычно наилучшим доступным решением является соответствие сложной эквивалентности (см. 8.3).

Для перечислительных классификационных систем может быть достаточно обеспечить соответствия для всех перечисленных классов. Для синтетических систем (включая фасетные системы) возможности полноценного сопоставления должны также предусматривать классы, которые не перечислены, но могут быть синтезированы в соответствии с правилами системы. Синтезированные классы представляют собой сложные понятия. Как правило, они требуют соответствий сложной эквивалентности подходящим понятиям тезауруса (см. также 13.2.2).

Хотя иерархическая организация классификации может оказаться полезной для уточнения содержания данного класса, сами отношения классов не требуется сопоставлять (см. определение «сопоставления» в 3.40, которое применимо только к понятиям и классам, но не к отношениям).

Утверждения о соответствии должны быть выражены с использованием кода для представления класса, а не с использованием его наименования или соответствующих статей указателя.

Дополнительные рекомендации и примеры, применимые ко многим таксономиям и системам, используемым в управлении записями, а также к системам классификации, можно найти в разделе 13.

18 Классификационные системы для управления документами

18.1 Ключевые характеристики и происхождение

18.1.1 Общее описание

Многие организации используют специально разработанные системы классификации для управления своими официальными документами (иногда их называют файловым планом или классификационной системой для бизнеса). Если системы классификации, используемые библиотеками (см. раздел 17), обычно поддерживают доступ в соответствии с темой документов, то схемы, используемые в системах управления документами, больше связаны с процессами бизнеса. Следовательно, каждая запись обычно связана с бизнес-функциями, операциями и транзакциями, которые она регистрирует в правильном хронологическом порядке. Хорошо разработанная система обычно основана на анализе деятельности организации, нормативно-правовой базы и факторов риска, а также требований безопасности и отчетности.

На рисунке 13 показана типичная функциональная система классификации, используемая для управления документами, и отражены только несколько наименований и идентификаторов классов высшего уровня. Все показанные классы, вероятно, могут быть неоднократно подразделены еще на несколько уровней.

Идентификатор	Наименование класса
HR000	Управление персоналом
HR100	Поддержание оптимального уровня кадрового обеспечения
HR110	Оценка кадровых потребностей
HR120	Набор кадров
HR121	Публикация объявлений
HR122	Отбор кандидатов
HR123	Собеседование
HR124	Оповещение кандидатов
HR200	Учет отпусков

Рисунок 13 — Выписка из функциональной классификационной системы

Классификационные системы для управления документами, как правило, отвечают юридическим требованиям (которые могут варьировать в зависимости от юрисдикции) к ведению полных и точных записей о деятельности организации. Требования к хранению и удалению являются ключевым фактором проектирования, потому что управление документами проще, если классы организованы так, что все содержимое отдельных файлов и папок может храниться в течение требуемого периода и утилизироваться в соответствии с заранее определенным графиком. На способ группировки файлов также может влиять необходимость управления доступом.

Примечание — Для получения дополнительной информации см. ***ГОСТ Р ИСО 15489-1.***

18.1.2 Место и роль в информационном поиске

Система классификации в системе управления документами обычно поддерживает просмотр документов и играет ограниченную роль в поиске.

В качестве вспомогательного средства просмотра иерархическая структура системы помогает пользователю находить взаимосвязанные серии записей на любом уровне агрегации. Например, пользователь может найти папку, содержащую всю переписку с американскими акционерами за определенный год, и просмотреть ее, чтобы найти подпапку только для тех акционеров, которые находятся в Калифорнии. В этой подпапке пользователь может выбрать файл для одного конкретного акционера и может просмотреть этот файл для идентификации одной записи — возможно, запроса, полученного от акционера.

Когда имена файлов и папок в иерархической структуре известны пользователю электронной системы управления документами, они обычно могут использоваться в поисковых запросах. Но этих имен обычно недостаточно при поиске записей по конкретной теме, поэтому необходимы дополнительные инструменты поиска, такие как полнотекстовый поиск и/или индексирование с помощью тезауруса.

18.1.3 Создание и развитие

В те дни, когда большинство записей производились на бумаге, отдельные документы обычно группировались в физические подшивки, которые собирались в папки и сохранялись в шкафах для хранения документов. Подшивки, папки и картотеки помечались в соответствии с системой классификации, которая могла быть или не быть полностью представлена отдельным набором таблиц. С появлением компьютеров такие системы были постепенно автоматизированы, так что в настоящее время электронные системы используются для подавляющего большинства записей. Все чаще эти системы приводятся в соответствие с [4], [5], и практика систем электронного управления документами быстро развивается. Однако аналогия с подшивками и папками вместе с традицией иерархической организации по-прежнему определяют классификационные системы большинства систем управления документами.

18.1.4 Словарный контроль

В системе управления документами для составления названий отдельных записей, а также для маркировки подшивок и папок, часто используют наименования классов классификационной системы. Различные системы применяют такой словарный контроль по-разному, но в пределах одной системы согласованность наименований имеет очень большое значение.

При соблюдении ГОСТ Р ИСО 15489-1 верхние уровни системы обычно представляют собой основные бизнес-функции организации, второй уровень отражает деятельность в пределах бизнес-функции, третий и последующие уровни являются дополнительными уточнениями действий и групп операций в рамках каждого вида деятельности.

Чтобы добиться уникальной идентификации каждого класса в системе, его имя комбинируют с именами вышестоящих классов, либо используют кодовый идентификатор или комбинацию этих способов, в зависимости от правил системы.

18.2 Семантические компоненты и отношения в сопоставлении с компонентами тезауруса

18.2.1 Имена классов

Каждый класс в системе обычно имеет имя или метку, сравнимые с наименованием в тематической классификационной системе (см. 17.2.2). В разных системах управления документами имена классов могут называться «рубрики», или «дескрипторы», или «ключевые слова», или «заголовки», или «метки». В каждом классификационном ряду каждое имя класса должно быть уникальным, чтобы пользователь не путался. Но в разных рядах одно и то же имя может повторяться. Например, «Бронирование» может встречаться как в разделе «Конференции», так и в разделе «Выставки» как название двух совершенно разных классов. В структуре, показанной на рисунке 13, имя класса «Публикация объявлений» может повториться и в контексте продажи недвижимости.

18.2.2 Идентификаторы

Системы электронного управления документами обычно присваивают уникальный идентификатор (такой, как структурированный цифровой или буквенно-цифровой ссылочный код) каждому электронному классу или файлу вплоть до уровня отдельных записей. Этот идентификатор сопоставим с кодом тематической классификационной системы, за исключением того, что он распространяется на документы, а не только на классы структуры.

Отдельный идентификатор не всегда присутствует в бумажных системах, если имена классов считаются достаточными для определения местоположения определенных записей. Но там, где он присутствует, он часто принимает форму мнемонического кода класса верхнего уровня, за которым следуют буквы или цифры, соответствующие каждому подразделению на последовательных уровнях. Например, идентификатор «REG-PUB-COR» может представлять подпапку «Переписка» (Correspondence) в папке «Общественное обсуждение» (Public consultation) класса верхнего уровня «Региональное планирование» (Regional planning). Отдельный документ в подпапке может иметь идентификатор «REG-PUB-COR 067».

18.2.3 Другие компоненты системы

С каждым классом в системе обычно связаны дополнительные данные. Часто встречаются следующие элементы:

- описание (сравнимое с лексическим примечанием тезауруса и с описанием класса тематической классификационной системы);
- индексные термины (включая соответствия терминам тезауруса);
- даты (открытия, закрытия или удаления файла);
- права доступа (с указанием, кому разрешен доступ к контенту);
- график хранения (с указанием срока хранения записей в этом файле);
- исторические примечания (сопоставимые с историческими примечаниями к понятиям в тезаурусе);
- перекрестные ссылки на другие классы в системе (сравнимые со ссылками «см. также» в тематической системе классификации и с ассоциативными отношениями в тезаурусе).

Для поддержки надлежащего применения системы в нее часто добавляется указатель. Запись в указателе может состоять из имени класса и его вышестоящих классов, а также может дополняться полезными терминами из описания или синонимами, а также индексными терминами, назначенными классу.

Кроме того, к таблицам прилагаются правила, позволяющие администраторам расширять систему, разделять или объединять классы в процессе ввода новых записей и серий записей.

18.2.4 Иерархическая структура

Как и любая классификационная система, система управления документами предполагает предкоординацию понятий (см. 13.1) и обычно является моноиерархической. Однако некоторые системы позволяют причислять документ более чем к одному классу или файлу в системе.

18.3 Рекомендации по совместимости с тезаурусом

Следует рассмотреть следующие возможные варианты использования:

а) тезаурус может быть использован при составлении указателя к таблицам системы. Хотя этот процесс индексирования можно рассматривать как сопоставление с тезаурусом, это не будет обсуж-

даться далее, поскольку получающиеся «соответствия» используются внутри системы (обычно для поддержки задачи ввода новых документов в систему), а не в целях внешнего взаимодействия;

b) тезаурус также может использоваться при подготовке предметного указателя документов, хранимых или контролируемых системой управления документами. Этот указатель является более обширным, чем описанный в a), поскольку индексируются отдельные документы. Тезаурус, используемый таким образом, дополняет систему классификации, позволяя осуществлять поиск на основе тематического содержания, а не бизнес-контекста, к которому адресует система классификации;

c) другой вариант использования возникает, когда документы из системы управления документами выбираются для включения в другую поисковую систему, в которой используется другой тезаурус. Стоимость индексирования по новому тезаурусу может быть уменьшена, если использовать соответствия от надлежащих классов системы классификации. Соответствия позволяют добавлять подходящие термины тезауруса в метаданные выбранных документов [см. также 12.1 a)];

d) вариант использования, описанный в b), можно альтернативно рассматривать как сопоставление терминов тезауруса с системой классификации и использование соответствий для преобразования запросов во время поиска [см. также 12.1 b)].

Из этих случаев только c) и d) включают сопоставление системы классификации и внешнего тезауруса. Общее руководство здесь дают рекомендации и примеры в разделе 13 с учетом следующих дополнительных указаний:

- прежде чем какие-либо соответствия будут установлены, следует определить наиболее подходящий способ представления каждого класса в утверждениях о соответствии, следя за тем, чтобы выбранный способ идентификации был уникальным в системе;
- соответствия в варианте использования c) направлены от классификационной системы к тезаурусу, без обязательной обратимости. Могут применяться типы соответствия, описанные в разделах 7—11. Однако кумулятивную сложную эквивалентность следует применять с осторожностью, поскольку это может отрицательно повлиять на точность поиска;
- соответствия в варианте использования d) направлены от тезауруса к системе классификации без обязательной обратимости. Могут применяться типы соответствия, описанные в разделах 7—11. Однако пересекающаяся сложная эквивалентность должна применяться только в том случае, если система является полииерархической, и/или назначение документу более одного класса является обычной практикой;
- в случаях c) и d) следует понимать, что как полнота, так и точность, которые можно достичь с помощью таких соответствий, будут хуже, чем то, что можно сделать, индексируя отдельные записи непосредственно с помощью внешнего тезауруса. Это связано с тем, что классы системы управления документами редко разрабатываются для отражения предметного содержания документов (см. 18.1.1). Для получения хороших результатов поиска более эффективен подход, заключающийся в индексировании отдельных документов, как в b). Если требуется доступ с помощью второго, отличного тезауруса, эффективное решение могут обеспечить соответствия между этими тезаурусами.

19 Таксономии

19.1 Ключевые характеристики и происхождение

19.1.1 Общее описание

Типичная таксономия представлена иерархическим словарем, используемым для классификации, распределения по категориям, организации, просмотра, навигации, поиска и/или фильтрации какого-либо контента в сетевой среде. Распространенным вариантом использования является поддержка навигации, особенно путем иерархической организации и просмотра широкого набора электронных ресурсов, например, сайтов, баз данных, порталов, вики-ресурсов. Таксономии часто используют для представления меню сайта. Чтобы дополнить навигационные функции возможностью поиска, таксономии могут включать синонимы, действующие как входные термины, и ссылки «см. также» между связанными категориями в иерархии.

Некоторые таксономии настроены так, чтобы отражать язык, культуру и цели конкретной организации, и используются в качестве основы для обмена знаниями внутри организации. Их можно рассматривать как карты знаний или как средства коммуникации и обучения, объединяющие историю, опыт и внутреннюю информацию для поддержки деловых операций.

Некоторые таксономии включают правила автоматической категоризации и даже персонализации входящих документов, например, новостных сообщений. Правила могут использоваться для сортировки входящих элементов по категориям и избирательной доставки их пользователям, проявляющим интерес к определенным категориям.

Примечание — Методы автоматической категоризации и персонализации выходят за рамки этого стандарта.

19.1.2 Место и роль в информационном поиске

Таксономии обычно используют для навигации по ресурсам в порталах, интрасетях и на вебсайтах, и они поддерживают поиск, главным образом, путем обеспечения просмотра.

Чтобы облегчить просмотр, категории в таксономии обычно располагают в иерархическом порядке, так что пользователь может перемещаться вверх или вниз до необходимого уровня. Ссылки между связанными категориями в разных иерархиях позволяют осуществлять боковой просмотр.

Для облегчения поиска категории часто помечены несколькими синонимами. Пользователь может найти то, что он хочет, введя любой из них.

Для облегчения фильтрации имеется возможность выдавать категории, в которых имеется поисковый термин, так что пользователь может сузить поиск до необходимых категорий.

При другом подходе к поиску простую таксономию связывают с более сложным тезаурусом. При этом таксономия выступает как простой иерархический набор категорий и подкатегорий. Однако за ним может находиться гораздо более сложный тезаурус. Если каждая категория в таксономии сопоставлена с соответствующим понятием тезауруса, все предпочтительные и непредпочтительные термины, а также любые лексические примечания и отношения, которыми обладает это понятие, могут использоваться для обеспечения дополнительных возможностей поиска.

19.1.3 Создание и развитие

Термин «таксономия» происходит из греческого языка: «таксис» («порядок» или «расположение») и «номос» («закон» или «наука»). Первоначально этот термин часто применяли для обозначения задачи классифицирования и именования организмов исключительно с моноиерархической точки зрения. В этом стандарте, однако, определение таксономии расширено, чтобы охватить как моноиерархические, так и полииерархические классификации любых предметов. В корпоративных системах таксономии обычно включают такие точки входа, как бизнес-функции, типы продукции, деловые отношения, отрасли хозяйства, бизнес-события, а также типы документов и записей.

19.1.4 Словарный контроль

Некоторые таксономии следуют модели систем классификации, предоставляя уникальные коды для каждой категории. Однако гораздо чаще метка категории не сопровождается классификационным кодом. Если таксономия следует модели тезауруса, каждая метка категории будет уникальной в рамках системы и может использоваться отдельно для обозначения конкретной категории. Альтернативно или дополнительно может быть предоставлен уникальный идентификатор, хотя, как правило, он скрыт и предназначен в основном для использования компьютером.

19.2 Типы таксономий

19.2.1 Общие положения

Термин «таксономия» используется (в том числе и неправильно) настолько широко, что это название может быть присвоено множеству самых разных типов словарей. Однако в этом разделе возможные типологии таксономий обсуждаться не будут, поскольку основное внимание уделяется таксономиям, используемым для навигации по ресурсам. Даже они включают такое разнообразие, что в контексте функциональной совместимости полезно различать некоторые ключевые структурные особенности, которые влияют на их сопоставление, как описано в 19.2.2 и 19.2.3.

19.2.2 Моноиерархические и полииерархические структуры

В целях совместимости и в особенности для установления соответствий полезно отличать таксономии, которые следуют модели систем классификации, от таксономий, разработанных скорее как тезаурусы. На рисунках 14 и 15 показаны примеры различий.

Образование	Здоровье	Оборона
Учреждения	Учреждения	Учреждения
Финансирование	Финансирование	Финансирование
Исследование	Исследование	Исследование

Рисунок 14 — Пример простой моноиерархической таксономии
(допускающей повторение меток категорий в разных областях)

Рисунок 14 аналогичен системе классификации, показанной на рисунке 8 в разделе 13, за исключением того, что здесь нет кодов. Наименования категорий повторяются в разных иерархиях, но эти категории различны. «Финансирование» в иерархии образования относится только к финансированию образовательной деятельности; «Финансирование» в иерархии здоровья — к финансированию мероприятий в области здравоохранения и так далее. Как и наименование класса в типичной классификационной системе, наименование категории недостаточно для однозначного определения категории и не должно использоваться отдельно в утверждениях о соответствии.

На рисунке 15 показана таксономия, в которой каждое наименование категории уникально, подобно предпочтительным терминам тезауруса. Одна и та же категория (например, «Военная подготовка») может появляться в нескольких иерархиях; но если это так, то объем и содержание категории одинаковы, независимо от ее местоположения. Когда применяются уникальные наименования категорий, их можно надежно использовать в утверждениях о соответствии.

Несмотря на некоторые сходства с тезаурусами, следует отметить, что на рисунке 15 правила иерархии не так строги, как у истинного тезауруса (см. ГОСТ Р 7.0.91—2015, подраздел 10.2). Например, иерархическая связь между «Образованием» и «Образовательными учреждениями» не рекомендуется для тезауруса.

Образование	Здоровье	Оборона
Образовательные учреждения	Медицинские учреждения	Оборонные учреждения
Образовательная деятельность	Деятельность, связанная с охраной здоровья	Оборонная деятельность
Обучение, подготовка	Медицинская подготовка	Военные операции
Медицинская подготовка	Превентивная медицина	Военная подготовка
Военная подготовка	Хирургические операции	и т. д.
Подготовка учителей	и т. д.	
и т. д.		

Рисунок 15 — Пример полииерархической таксономии
(одна и та же категория может быть включена в несколько иерархий)

19.2.3 Несимметричные структуры в многоязычных и мультикультурных таксономиях

Очень много таксономий разработано для специфических сообществ, где их используют неподготовленные лица. Таксономии подвергаются культурным и социальным влияниям, которые различаются в параллельных сообществах, где говорят на разных языках. *По этой причине в многоязычных таксономиях обычно встречаются несимметричные структуры (описанные в ГОСТ Р 7.0.91—2015, пункт 10.2.6).*

19.3 Семантические компоненты и отношения

19.3.1 Категории

Основные единицы таксономии, соответствующие понятиям тезауруса или классам системы классификации, обычно называются «категориями». Категория может охватывать одно понятие, например «Права человека» или комбинацию понятий, например «Правительство, граждане и права». Подобная комбинация часто создается как инструмент представления группировки нескольких более конкретных категорий. Иногда комбинация понятий в категории лучше всего передается с помощью фразы, например «Борьба с раком» или «Поиск для клинических испытаний?», соответствующей ожидаемым потребностям пользователей. Иногда значение категории, отдельного понятия или сочетания

понятий выражается императивно, чтобы убедительно передать смысл (например, «Будь осторожен на солнце» или «Держись подальше от табака»). На более низких уровнях таксономии часто встречаются предкоординированные выражения, такие как «Безопасность автомобиля и детские сиденья» или «Сообщение о преступности и антиобщественном поведении».

В дополнение к категориям, относящимся к предмету или теме документов, таксономии иногда включают нетематические категории, такие как:

- предполагаемая аудитория (например, «Дети и подростки»);
- ограничение по времени или месту (например, «Архив»);
- форма документов (например, «Пресс-релизы»);
- связанная задача (например, «Претензии»).

На любом уровне таксономии может иметь место любое сочетание тематических категорий с указанными выше типами.

Каждой категории присваивается метка, которая в одних таксономиях уникальна, но в других неоднозначна (см. рисунки 14 и 15). *Там, где требуется устранение неоднозначности, метке иногда присваивают релятор (см. ГОСТ Р 7.0.91—2015, пункт 6.2.2).* Метки категорий соответствуют предпочтительным терминам в тезаурусе. В многоязычной таксономии каждая категория обычно имеет различные метки для разных языков.

В некоторых таксономиях каждая категория помимо метки, применяемой человеком, имеет уникальный несемантический идентификатор для машинного использования. Это особенно полезно, когда ярлыки категорий неоднозначны. Обеспечение удобочитаемости кодов — редкость.

19.3.2 Варианты синонимов

Категориальные метки в таксономиях часто имеют синонимы подобно тому, как дескрипторы в тезаурусе имеют аскрипторы. Как и термины в синонимическом ряду, они могут включать неточные синонимы, аббревиатуры, сокращения и лексические варианты (см. 24.2). Для устранения неоднозначности к терминам иногда добавляют реляторы. Отношение эквивалентности между меткой категории и ее синонимами может отображаться явно, но чаще всего оно скрыто. Альтернативный способ вызова синонимов состоит в том, чтобы установить соответствия категорий с надлежащими единицами тезауруса или набора синонимических рядов. Тогда их можно использовать, вызывая соответствия во время поиска, хотя синонимы и не присутствуют в самой таксономии.

19.3.3 Иерархические отношения

В некоторых таксономиях соблюдаются правила, подобные тем, что регулируют иерархические отношения в тезаурусах (см. ГОСТ Р 7.0.91—2015, подраздел 10.2), включая использование полииерархических структур (см. 19.2.2). Однако чаще всего иерархии в таксономии похожи на иерархические классификационные системы (см. 17.2.3), и их структура обычно моноиерархическая (см. 19.2.2) с неявной предкоординацией понятий. См. раздел 13 для более подробного рассмотрения того, как управлять предкоординацией.

19.3.4 Ассоциативные отношения

Некоторые таксономии включают в себя ассоциативные отношения между смежными категориями, обычно из разных иерархий, сравнимые с ассоциативными отношениями в тезаурусе (см. ГОСТ Р 7.0.91—2015, подраздел 10.3). Они могут отображаться как ссылки «см. также», как во многих системах классификации, и/или быть реализованы как гиперссылки между связанными категориями.

19.3.5 Определения и лексические примечания

К категории, где это необходимо, для уточнения ее содержания могут прилагаться определения и/или лексические примечания.

19.3.6 Правила автоматической категоризации

Факультативно к категориям могут быть добавлены правила автоматической категоризации, автоматического опроса и персонализации.

19.4 Сопоставление тезауруса и таксономии

19.4.1 Общие положения

Обычно термин «таксономия» применяется широко, с большим разнообразием значений. Поэтому, прежде чем будут установлены соответствия для словаря, описанного как «таксономия», необходимо его изучить и определить его тип и основные характеристики. Характеристики, которые могут влиять на содержание и форму представления соответствий, включают:

- наличие/отсутствие предкоординации (см. раздел 13, особенно рисунок 8);

- наличие/отсутствие полииерархии;
- наличие/отсутствие несимметричных структур (в многоязычной таксономии);
- наличие кодов категорий;
- наличие уникальных категориальных меток, функционирующих как дескрипторы;
- наличие имен собственных.

Если обнаружена предкоординация в стиле рисунка 8, таксономия может рассматриваться как тип классификационной системы, и следует соблюдать рекомендации, изложенные в разделе 13 и 17.3. Однако, если каждая категориальная метка уникальна в таксономии, и особенно там, где разрешены полииерархические структуры, таксономия может рассматриваться скорее как тезаурус, и предкоординация подчиненных и вышестоящих категорий с меньшей вероятностью вызовет трудности.

В некоторых задачах требуется установление соответствий единиц тезауруса единицам таксономии, а в других — соответствий в противоположном направлении. Возможен любой из этих вариантов, и примеры в разделах 12 и 13 могут помочь при выборе.

Могут применяться все типы соответствий, описанные в разделах 7—11.

Соответствие сложной эквивалентности обычно требуется при сопоставлении объединительной категории, такой как «Правительство, граждане и права» или «Аварии и предотвращение несчастных случаев».

Во всех утверждениях о соответствии категория должна быть однозначно обозначена с использованием категориальной метки, если она является уникальной, или несемантического уникального идентификатора. Первый вариант, как правило, больше подходит для читателей; последний — для операций на компьютере.

В случае несимметричной многоязычной таксономии следует сделать выбор относительно того, какую языковую версию или версии использовать в соответствиях. В симметричных частях таксономии соответствие с любой категорией должно быть одинаково действительным на всех ее языках. В несимметричных частях для достижения полного соответствия каждая категория в каждой языковой версии должна отображаться отдельно. В некоторых контекстах может быть достаточно подготовить соответствия только для одной из языковых версий. Но в этом случае соответствия не должны применяться к контекстам, использующим другие языки.

Для нетематических категорий сложно находить эффективные соответствия, поскольку тезаурусы, предназначенные для предметного индексирования, обычно не содержат терминов или понятий, необходимых для индексирования других элементов метаданных, таких как тип документа или аудитория. Если тезаурус не используется в качестве источника значений для этих элементов, соответствий между таксономией и тезаурусом устанавливать не следует. Для конкретного приложения иногда возможно разработать соответствия от нетематических категорий к комбинации терминов, взятых частично из тезауруса и частично из нормативного файла, используемого для рассматриваемого элемента метаданных (см. пример 7 в 19.4.2).

19.4.2 Практические примеры

В следующих примерах во всех формулах обозначения «ТЕЗ» — это идентификатор тезауруса, а «ТАКС» — идентификатор таксономии. В этих примерах предполагается, что метка категории уникальна; а если это не так, то в формулировках соответствия ее будет заменять какой-то другой однозначный идентификатор.

Примеры:

1 Описание соответствия:

В таксономии есть категория «права», а ближайший термин в тезаурусе — «гражданские права».

Формулировка соответствия:

права ТЕЗ УС гражданские права

Обсуждение:

В этом случае отображение из таксономии в тезаурус почти такое же, как между двумя тезаурусами. См. разделы 7—10.

Соответствие в обратном направлении:

гражданские права ТАКС ШС права

2 Описание соответствия:

В таксономии есть категория «права человека», а ближайший термин в тезаурусе — «гражданские права».

Формулировка соответствия:
права человека ТЕЗ ~ЭК гражданские права
или
права человека ТЕЗ УС гражданские права
Обсуждение:
Какое из этих соответствий является более точным, зависит от контекста, в котором соответствия будут использоваться; например, от того, в соответствии с какой юрисдикцией определяются права.
Соответствие в обратном направлении:
гражданские права ТАКС ~ЭК права человека
или
гражданские права ТАКС ШС права человека

3 Описание соответствия:
Таксономия имеет категорию «джемы, желе и консервы», а ближайшее понятие в тезаурусе — «фруктовые консервы».
Формулировка соответствия:
джемы, желе и консервы ТЕЗ ЭК фруктовые консервы
Обсуждение:
Возможно, формулировку соответствия можно изменить, чтобы отметить его точность или неточность. Чтобы определить, какой вариант подходит, необходимо изучить все примечания к понятиям, иерархическое окружение категории и способы отнесения документов к категориям и понятиям.
Соответствие в обратном направлении:
фруктовые консервы ТАКС ЭК джемы, желе и консервы

4 Описание соответствия:
Таксономия имеет категорию «дети и ожирение», в то время как тезаурус имеет отдельные понятия «дети» и «ожирение».
Формулировка соответствия:
дети и ожирение ТЕЗ ЭК дети + ожирение
Обсуждение:
Хотя логическая связка «И» делает этот случай похожим на пример 3, на самом деле это пример пересекающейся сложной эквивалентности, и с ним нужно обращаться по-другому. Эта категория не предназначена для группировки всего, что касается детей, и всего, что касается ожирения, а только для тематики, связанной с сочетанием понятий.
Соответствие в обратном направлении:
Должны быть рассмотрены следующие формулировки:
дети ТАКС УС дети и ожирение
ожирение ТАКС УС дети и ожирение
Перед тем, как принять любую из них, таксономию нужно просмотреть для поиска других категорий, которые отображаются в тезаурус более точно или так же точно. Может быть установлено несколько соответствий к одному понятию.

5 Описание соответствия:
В таксономии есть категория «правительство, граждане и права», а в тезаурусе нет единого понятия, объединяющего все эти темы. Тем не менее, в тезаурусе есть отдельные понятия «правительство», «общественность» и «гражданские права», а также несколько других понятий, связанных с правами.
Варианты формулировки соответствия:
a) правительство, граждане и права ТЕЗ ЭК правительство | общественность | гражданские права
b) правительство, граждане и права ТЕЗ ЭК (правительство | гражданские права) + общественность
c) правительство, граждане и права ТЕЗ ЭК правительство | гражданские права
d) правительство, граждане и права ТЕЗ ЭК правительство | гражданские права | защита потребителя
e) [нет соответствия]
Обсуждение:
На первый взгляд, это выглядит как пример кумулятивной сложной эквивалентности, как показано в варианте a).
Тем не менее, вариант a) вводит в заблуждение, потому что эта категория на самом деле не включает информацию о гражданах или общественности. Упоминание граждан предназначено для того, чтобы предположить, что информация предназначена для граждан (а не для предприятий или

государства). Вариант b) кажется ближе, потому что он несет информацию о правительстве и/или о гражданских правах, в которой также упоминается об общественности.

Слабость варианта b), если он применяется к поисковому предписанию, заключается в необходимости отражения «общественности» в метаданных искомых ресурсов. Даже когда документ написан специально для руководства гражданам и/или широкой общественности, этот аспект обычно не отражается в его метаданных. Чтобы избежать исключения релевантного материала, лучше использовать вариант c) (хотя такой расширенный запрос может не дать хорошей точности).

Вариант d) может быть результатом более тщательного изучения таксономии и выявления подкатегории «права потребителей» среди подкатегорий категории «правительство, граждане и права». Права потребителей обычно не рассматриваются в ряду гражданских прав, и поэтому в формулу соответствия необходимо добавить больше понятий, что сделает ее более расплывчатой, чем раньше.

Очевидно, что все предложенные соответствия неточны, но маркер неточности не нужен, потому что все соответствия сложной эквивалентности неточны. В данном случае соответствия кажутся настолько неточными, что возникает вопрос, действительно ли необходимо установление соответствий.

Вариант e) представляется целесообразным, если использование соответствия несущественно для пользователей. Очевидно, что категория «правительство, граждане и права» полезна в таксономии как инструмент, помогающий пользователям переходить к подкатегориям, таким как «гражданство», «удостоверения личности», «права потребителя» и т. п., но менее полезна в поисковых предписаниях или в метаданных документов. Соответствие с обобщающей категорией, подобной этой, часто можно игнорировать, если выполнить тщательную работу по установлению соответствий всех ее подкатегорий.

Соответствие в обратном направлении:

Единственное обратное соответствие установить невозможно. Должны быть рассмотрены следующие соответствия:

a) правительство ТАКС ШС правительство, граждане и права

b) общественность ТАКС ШС правительство, граждане и права

c) гражданские права ТАКС ШС правительство, граждане и права

d) защита потребителя ТАКС ШС правительство, граждане и права

Вариант g) неуместен, если категория не связана соответствием с «граждане» и «общественность».

Могут быть справедливы варианты f), h) и j), но прежде чем принять любой вариант, следует проверить подкатегории в категории «правительство, граждане и права». Среди них могут иметься такие категории (как «права потребителей»), которые намного точнее соответствуют понятиям тезауруса, чем более широкая категория. Соответствия на одном уровне специфичности обычно предпочтительнее, чем иерархические соответствия, и поэтому, например, формулировка

защита потребителя ТАКС ~ЭК права потребителей предпочтительнее формулировки

защита потребителя ТАКС ШС правительство, граждане и права.

6 Описание соответствия:

В таксономии есть категория «Преодоление рака», а в тезаурусе нет единого понятия, объединяющего «преодоление» и «рак». Тезаурус имеет понятие «рак». Но ближайший отдельный термин/понятие к «преодоление» — это «управление болезнью», «лечение» или «охрана здоровья», и ни один из них, похоже, не отражает того, что подразумевается в таксономии. Категория «Преодоление рака» актуальна для сообщества больных раком, лиц, осуществляющих уход, и специалистов, но она состоит из очень многих аспектов, таких как терапия, паллиативная помощь, консультирование, уход за детьми, лекарства, образ жизни и т. п.

Варианты формулировки соответствия:

a) борьба с раком ТЕЗ ЭК рак + (услуги поддержки | лечение | дополнительная медицина | паллиативная помощь | консультирование | уход за детьми | лекарства | образ жизни | временное облегчение)

b) борьба с раком ТЕЗ АС рак

Обсуждение:

Вариант a) стремится охватить все то, что подразумевается в «преодолении», но результат далеко не идеален. Подобная формулировка не может использоваться для преобразования индексных терминов (кроме как в системе подсказок для индексаторов). При использовании для преобразования поисковых предписаний нельзя ожидать высоких показателей полноты и точности.

Вариант b) признает сложность охвата всей сущности «преодоления». Опять же, это может быть полезно в качестве подсказки для пользователя, осуществляющего поиск, или индексатора, но не подходит для автоматического преобразования поисковых предписаний или индексных терминов.

Соответствие в обратном направлении:

Должна быть рассмотрена следующая формулировка:

рак ТАКС АС преодоление рака

Перед принятием этого соответствия необходимо проверить, имеет ли таксономия более узкую категорию, более точно соответствующую понятию «рак». Если она имеется, то эквивалентное соответствие предпочтительнее иерархических или ассоциативных соответствий.

Могут также рассматриваться следующие формулировки:

служба поддержки ТАКС АС преодоление рака

лечение ТАКС АС преодоление рака

и т. п.

В каждом случае таксономию следует проверять на наличие других, более близко совпадающих категорий, и предпочтительной должна быть наиболее близкая.

7 Описание соответствия:

Таксономия для сайта, посвященного здравоохранению, имеет категорию наивысшего уровня «Дети и подростки», а в тезаурусе есть понятия «дети» и «подростки», а также «здравоохранение». Но в системах и коллекциях, применяющих этот тезаурус, источником терминов, используемых для элемента метаданных «Аудитория», является небольшой нормативный список под названием «ЦЕЛИ», который содержит значения «Дети» и «Молодежь».

Варианты формулировки соответствия:

a) Дети и подростки ТЕЗ ШС здравоохранение

b) Дети и подростки ЦЕЛИ ШС дети \| молодежь

c) Дети и подростки ЭК ЦЕЛИ (дети \| молодежь) + ТЕЗ здравоохранение

Обсуждение:

Категория таксономии, помеченная как «Дети и подростки», имеет содержание более узкое, чем буквальное значение этой метки. Оно фактически является здравоохранением для детей и подростков. Эта метка вводит в заблуждение, когда вырвана из контекста.

Соответствие этой категории понятиям тезауруса «дети» и «подростки» неуместно, поскольку оно может привести к выдаче документов о детях, а не документов, написанных для детей. Оно также вводило бы в заблуждение, если применялось бы к преобразованию индексных терминов.

Вместо того, чтобы вообще не устанавливать никакого соответствия, возможен вариант a), хотя он не определяет целевую аудиторию. Следует использовать функцию ШС (а не ЭК), поскольку «здравоохранение» — это более широкое понятие, применимое ко всему контенту сайта.

Вариант b) действителен только в том случае, когда приложение предусматривает поиск (или преобразование индексных терминов) в поле метаданных «Аудитория», а не в поле «Тема». Он также использует метку ШС, а не ЭК, чтобы при поиске выдавались все материалы, предназначенные для молодой аудитории, а не только имеющие отношение к здравоохранению.

Вариант c) является потенциально более мощным, объединяя желаемую предметную область с ограничением по аудитории. Но используемый им синтаксис не является стандартным для внешних поисковых систем. Результаты, скорее всего, будут эффективны только в специализированном приложении, контролируемом администраторами системы.

Соответствие в обратном направлении:

С применением любого из этих вариантов в обратном направлении связаны определенные риски. Самое большее, их можно использовать в интерфейсе для подсказки, когда предлагаются варианты терминов при индексировании или поиске.

20 Словари предметных рубрик

20.1 Ключевые характеристики и происхождение

20.1.1 Общее описание

Словари предметных рубрик — это такой тип контролируемого словаря, который используется для представления в синтезированной форме тем, обсуждаемых в документах любого типа. Словари предметных рубрик имеют некоторые общие характеристики с тезаурусами и системами классификации. Подобно тезаурусу, они представляют понятия в форме терминов или фраз, а подобно синтетической системе классификации, они предоставляют синтаксические правила для объединения терминов в предкоординированные строки, которые представляют более сложные понятия и темы. Хорошо известными примерами словарей предметных рубрик являются «Предметные рубрики Библиотеки Конгресса» (LCSH) [9], «Медицинские предметные рубрики» (MeSH) [24], ***«Предметные рубрики Российской национальной библиотеки (ПР РНБ)».***

20.1.2 Место и роль в информационном поиске

Основной функцией словарей предметных рубрик является представление связанных тем, позволяющих упорядочивать коллекции информационных ресурсов на основе их предметного содержания, а также облегчение просмотра и навигации по предметной области.

При индексировании индексатор отвечает за подбор всех компонентов, необходимых для представления всех аспектов темы, создавая тем самым сложную предметную рубрику. При поиске, когда программное обеспечение предлагает разные возможности, конечный пользователь может извлекать информационные ресурсы, используя любое отдельное слово, входящее в предметную рубрику, любую комбинацию слов, входящих в предметную рубрику, или предметную рубрику в целом. Использование отдельных слов и комбинации слов способствует полноте поиска, в то время как высокая точность достигается при использовании в качестве поискового запроса предметной рубрики в целом.

20.1.3 Создание и развитие

Словари предметных рубрик были созданы в конце 19-го века как инструмент, используемый для организации предметного доступа к информационным ресурсам, когда каталогизаторы начали формировать списки предметных терминов для обеспечения согласованности индексирования документов в каталоге своего учреждения. Начиная с этого времени, составлялись и использовались различные списки предметных рубрик. Предметные рубрики Библиотеки Конгресса (LCSH) являются наиболее широко используемой системой предметных рубрик не только в англоязычном мире, но также, благодаря переводам и адаптациям, в большом количестве учреждений, рабочий язык которых не является английским. *В России наиболее представительный словарь предметных рубрик, составляющий основу Национального авторитетного файла предметных рубрик, имеет Российская национальная библиотека.*

20.1.4 Словарный контроль

Подобно тому, как тезаурус обеспечивает представление данного понятия при индексировании и поиске одним и тем же термином (терминами), аналогичным образом словарь предметных рубрик способствует использованию для представления одного и того же понятия одних и тех же рубрик. Этим он отличается от системы классификации, которая применяет для этой цели классификационные коды.

20.1.5 Типы словарей предметных рубрик

Энциклопедические или политематические словари предметных рубрик охватывают все дисциплины и чаще всего используются для индексирования и поиска информационных ресурсов в политематических коллекциях. Словари *Предметные рубрики Библиотеки Конгресса (LCSH)* [23], *Справочник предметных рубрик Университета Лаваля (Répertoire des vedettes-matière de l'Université Laval, RVM)* [28], *Унифицированные справочник авторитетных предметных рубрик (RAMEAU)* [26], *Новый словарь для предметного индексирования (Nuovo Soggettario)* [10] и *Авторитетный файл ключевых слов (Schlagwortnormdatei, SWD),* который теперь входит в *Общий авторитетный файл (Gemeinsame Normdatei, GND)* [29], являются примерами политематических словарей предметных рубрик. Дисциплинарные или специализированные словари охватывают отдельную тему или дисциплину (например, *Медицинские предметные рубрики* (MeSH) [24] и *Предметные рубрики для музыки* (SHM) [12]), или конкретную категорию информационных ресурсов (например, *Тезаурус графических материалов* (TGM) [22]). Некоторые словари предназначены для конкретной категории конечных пользователей (например, *Предметные рубрики для детей Библиотеки Конгресса* [34]).

20.2 Семантические компоненты и отношения

20.2.1 Обзор

Основными компонентами словарей предметных рубрик являются рубрики, подрубрики и отношения между ними. Словарь может также включать или сопровождаться правилами комбинирования этих компонентов. *Возможен вариант, когда в словаре предметных рубрик содержатся уже сформированные сложные рубрики.*

20.2.2 Рубрики

Рубрики могут состоять из слова, представляющего одно понятие (например, «Арифметика»), из нескольких слов, представляющих одно понятие (например, «Эллиптическая биаксиальная геометрия»), или из нескольких слов, представляющих комбинацию различных понятий (например, «Шахты и минеральные ресурсы»). Форма и значение каждой рубрики контролируются, хотя и не так строго, как в тезаурусе. Например, в английских версиях словарей предметных рубрик обычно предпочитается множественное число счетных существительных (например, «Trees»; «Forests»; «Arid regions»). В многословных рубриках слова могут быть представлены в прямом естественном порядке (например, «Dacians in literature»; «River sardine fisheries») или в обратном порядке (например, «Civilization, Dacian»; «Series, Arithmetic»). Омографы дополняются реляторами, например, «Меркурий (планета)» и «Меркурий (римское божество)».

Предметные рубрики могут быть предназначенными для индексирования (используемые рубрики) или непредназначенными (отсылочные рубрики). Они сопоставимы с предпочтительными и непредпочтительными терминами, соответственно, в тезаурусе.

Рубрика может быть простой, состоящей только из одного понятия, или может быть сложной, если одна или несколько подрубрик (подзаголовков) добавлены к исходной рубрике (заголовку) для формирования предкоординированной строки.

20.2.3 Подрубрики

Сложные рубрики — это предкоординированные строки, созданные путем добавления к исходной рубрике (заголовку) одной или нескольких подрубрик, чтобы полно и точно представить тему. Двойное тире обычно соединяет подрубрику (подзаголовок) с исходной рубрикой (заголовком), как показано на рисунке 16. Подрубрики (подзаголовки) могут состоять из одного слова, например «Маркетинг», или из нескольких слов, например «Период Мэйдзи, 1868—1912». Функция подрубрик (подзаголовков) состоит в том, чтобы указать перспективу, точку зрения, форму представления и т. п. того, как тема, выражаемая рубрикой (заголовком), представлена в информационном ресурсе.

Типы подрубрик (подзаголовков) включают:

a) тематические подрубрики (подзаголовки), например «Маркетинг», «История», «Альтернативные методы лечения», «Колонии», «Генетические аспекты»;

b) географические подрубрики (подзаголовки), например «Италия», «Россия», «Квебек (провинция)»;

c) хронологические (временны́е) подрубрики (подзаголовки), например, «500—1400», «20-й век», «период Мэйдзи, 1868—1912»;

d) подрубрики (подзаголовки) формы документов, например, «Словари», «Справочники», «Статистика», «Руководства для любителей».

Многие подрубрики (подзаголовки) должны использоваться только со специфичными рубриками (заголовками) или с определенными категориями рубрик (заголовков), например, «Хранение — Болезни и повреждения: Использовать как тематический подраздел для отдельных растений и групп растений».

Небольшое количество подрубрик (подзаголовков), известных как свободно присоединяемые подрубрики (подзаголовки), могут быть добавлены к любой или к большинству рубрик (заголовков) словаря, например, «Живописные произведения: Использовать как формальный подраздел для названий стран, городов и т. п., имен отдельных персон, семей, наименований организаций и других сущностей, таких как парки, сооружения и т. п., а также для классов лиц, этнических групп, наименований войн и тематических рубрик (заголовков)».

Когда к рубрике (заголовку) добавляется более одной подрубрики (подзаголовка), они обычно отображаются в определенном порядке, например: тематическая — географическая — хронологическая — формы документов.

20.2.4 Отношения рубрик

Три основных типа отношений, встречающиеся в тезаурусах, используются также для структурирования большинства словарей предметных рубрик. Такими отношениями являются:

a) отношения эквивалентности (обозначенные перекрестными ссылками, такими как «Смотри», «Ссылка от», или ***метками USE и UF — в английском языке, «см» и «с» — в русском языке***);

b) иерархические отношения (обозначенные уровнем отступа, типографским способом, перекрестными ссылками, такими как «Смотри также более широкое понятие», «Смотри также более узкое понятие», или ***метками BT и NT — в английском языке, «в» и «н» — в русском языке***);

c) ассоциативные отношения (обозначенные перекрестными ссылками, такими как «Смотри также», или ***меткой RT — в английском языке, «а» — в русском языке***).

Характер и функции этих отношений в словарях предметных рубрик описаны в ГОСТ Р 7.0.91—2015, разделы 10, 11.

20.3 Сопоставление предметных рубрик с тезаурусными понятиями

Если в словаре предметных рубрик есть только простые рубрики, без каких-либо предкоординированных строк, сопоставление такой системы и тезауруса должно следовать тем же правилам, что и сопоставление двух тезаурусов (см. разделы 7—11). Однако при обработке сложных рубрик необходимы дополнительные указания.

Для иллюстрации положений этого раздела мы будем использовать примеры, взятые из словаря предметных рубрик, включающего рубрики и строки, показанные на рисунке 16.

автомобили
автомобили — справочники
автомобили — сцепления
автомобили — сцепления — справочники
автомобили — сцепления — техническое обслуживание — справочники
автомобили — техническое обслуживание
автомобили — техническое обслуживание — справочники
автомобили — тормоза
стиральные машины — техническое обслуживание — справочники
сцепления — техническое обслуживание
техническое обслуживание
тормоза

Рисунок 16 — Некоторые рубрики из словаря предметных рубрик

20.3.1 Сопоставление словаря предметных рубрик с тезаурусом

При сопоставлении словаря предметных рубрик с тезаурусом следует использовать три подхода. Подход должен быть выбран в контексте приложения с учетом следующих рекомендаций:

a) сопоставление с целевым тезаурусом отдельно каждой простой рубрики и каждой рубрики (заголовка) и подрубрики (подзаголовка) сложной рубрики в исходном словаре.

Пример — Предметная рубрика «автомобили —— техническое обслуживание» не отображается целиком, но «автомобили» и «техническое обслуживание» сопоставляются с целевым тезаурусом по отдельности.

Такой подход обеспечивает последовательное и всестороннее соответствие элементарных понятий, не требуя принятия во внимание синтаксических правил словаря предметных рубрик. Поскольку эти правила могут быть сложными, их нелегко применять в автоматизированной системе. Однако при этом подходе сложные понятия, выраженные предкоординированными строками в исходном словаре предметных рубрик, не получают соответствий, даже если в целевом тезаурусе имеется подходящее комплексное понятие. Это может привести к потере точности при поиске, если соответствия используются для преобразования индексных терминов и строк.

Пример — Несмотря на то, что сложная рубрика «автомобили — техническое обслуживание» не имеет прямого соответствия, но в качестве индексного термина эту рубрику можно сопоставить с комбинацией «автомобили» и «техническое обслуживание». Однако это неудовлетворительно, если целевой тезаурус включает в себя комплексное понятие «техническое обслуживание автомобилей». Пользователь, выполняющий поиск по этому термину тезауруса, не получит релевантные документы, заиндексированные сложной рубрикой;

b) сопоставление (дополнительно к процедуре a)) всех сложных рубрик, перечисленных в исходном словаре предметных рубрик. Если возможно, сопоставьте каждую из них с предкоординированным понятием в целевом тезаурусе. Если предкоординированное понятие в целевом тезаурусе не обнаруживается, сопоставьте исходную строку с комбинацией целевых понятий. Дальнейшие указания по сложной эквивалентности см. в 8.3.

Пример — Предметная рубрика «автомобили — техническое обслуживание» сопоставляется с целевым понятием «техническое обслуживание автомобилей», если оно существует в тезаурусе. Если его в тезаурусе нет, эта предкоординированная строка может быть сопоставлена комбинации «автомобили + обслуживание», как описано в 8.3.2.

Это расширяет число сопоставленных сложных понятий, но может привести к несоответствиям и неопределенности. Словарь предметных рубрик может включать строку «автомобили — техническое обслуживание» с правилом, что тематическую подрубрику (подзаголовок) «техническое обслуживание» можно использовать с любой другой рубрикой (заголовком), представляющей тип наземного транспортного средства. Следовательно, строка «велосипеды — техническое обслуживание», хотя и является действующей, не будет сопоставлена, поскольку только перечисленные строки сопоставлялись с целевым тезаурусом.

Ограничения как a), так и b) трудно преодолеть без вмешательства человека для уточнений при использовании соответствий, как для преобразования метаданных, так и для преобразования поисковых предписаний;

c) сопоставление именно тех предметных рубрик, которые фактически использованы для индексирования набора информационных ресурсов, а не всего словаря предметных рубрик. При этом будут найдены соответствия предкоординированным строкам, не перечисленным в исходном словаре, но в то же время исключены перечисленные строки, которые еще не использовались в качестве терминов индексирования в локальной коллекции ресурсов.

Этот подход имеет следующие преимущества:

- не нужно выделять людские и финансовые ресурсы для сопоставления предметных рубрик (заголовков) и подрубрик (подзаголовков), которые вряд ли когда-либо будут использоваться для индексирования;
- тот факт, что сопоставляются с целевым тезаурусом те и только те предметные рубрики, которые фактически использовались для индексирования, повышает точность и полноту поиска.

Его недостатки следующие:

- поскольку сложные предметные рубрики, построенные индексаторами, будут включать в себя множество одинаковых составных элементов в разных комбинациях, возрастает объем трудозатрат, и существует опасность предоставления конфликтующих соответствий для одних и тех же терминов (следует избегать повторного сопоставления одной и той же составляющей);
- результат процесса сопоставления действителен только для определенного каталога, указателя или базы данных и может быть недействительным для других каталогов, указателей или баз данных, даже если они используют тот же исходный словарь предметных рубрик в качестве словаря индексирования.

После того как с помощью какого-либо из вышеуказанных подходов будет подготовлен полный набор соответствий, его можно будет использовать либо для преобразования индексных терминов, либо для преобразования поисковых терминов (см. более подробно 12.1). Преобразование индексных терминов позволяет пользователю тезауруса получить доступ к коллекции, заиндексированной с помощью словаря предметных рубрик. Преобразование поисковых терминов позволяет пользователю словаря предметных рубрик получить доступ к коллекции, заиндексированной с помощью тезауруса.

20.3.2 Сопоставление тезауруса со словарем предметных рубрик

При отображении тезауруса в словарь предметных рубрик рекомендуемый подход заключается в сопоставлении всех понятий тезауруса по указаниям, изложенным в разделах 7—12. Если ресурсов недостаточно, экономия может быть достигнута путем сопоставления только тех понятий, которые на самом деле использованы в качестве индексных терминов в определенном каталоге, указателе или базе данных. Это ограничение, однако, вызовет проблемы, когда в коллекцию будут добавлены и проиндексированы документы по новой тематике с помощью ранее не использовавшихся терминов.

Все сложные понятия в исходном тезаурусе должны быть сопоставлены с имеющимися перечисленными рубриками словаря предметных рубрик. Если подходящая предкоординированная строка не перечислена в целевом словаре предметных рубрик, она должна быть создана путем объединения соответствующей рубрики (заголовка) с одной или несколькими подрубриками (подзаголовками) в соответствии с синтаксическими правилами, приведенными в словаре или в сопроводительной документации.

Пример — Исходное понятие «автомобилестроение» может быть сопоставлено с предметной рубрикой «автомобили — проектирование и производство», если правила целевого словаря предметных рубрик разрешают создание этой предкоординированной строки.

Когда подготовлен полный набор соответствий, он может быть использован для преобразования либо индексных, либо поисковых терминов (см. 12.1). Преобразование индексных терминов открывает пользователю словаря предметных рубрик доступ к коллекциям, заиндексированным по тезаурусу. Преобразование поисковых терминов открывает пользователю тезауруса доступ к коллекции, заиндексированной с помощью словаря предметных рубрик.

Автоматическое синтезирование предметных рубрик на основе совместной встречаемости терминов в метаданных определенных документов не рекомендуется без одобрения экспертом или специально разработанным алгоритмом. Например, если документу присвоены термины «справочники», «обучение», «автомобили», «сцепления», «коробки передач», «чистка», и «цены», то он может быть отнесен по крайней мере к семи рубрикам из числа рубрик, приведенных на рисунке 16, и ко многим другим. Без изучения документа трудно решить, какие из этих комбинаций наиболее пригодны. В этом

случае сопоставления следует устанавливать для понятий, действительно представленных в тезаурусе, но не для их комбинаций.

Пример — Если тезаурусным понятиям «техническое обслуживание автомобилей» и «цены» сопоставлены рубрики «автомобили — техническое обслуживание» и «цены» соответственно, то поисковое предписание «техническое обслуживание автомобилей И цены» может быть преобразовано в «(автомобили — техническое обслуживание) И цены».

В другом случае, когда индексные термины «техническое обслуживание автомобилей» и «цены» встретились в метаданных документа, они могут быть конвертированы в предметные рубрики «автомобили — техническое обслуживание» и «цены», но без изучения документа не следует создавать комбинированные индексные строки типа «автомобили — техническое обслуживание — цены».

21 Онтологии

21.1 Ключевые характеристики и происхождение

21.1.1 Общее описание

Термин «онтология» имеет много разных значений, начиная от семантических моделей данных и заканчивая философским применением категориального и метафизического анализа понятий предметной области. В компьютерных науках, более конкретно — в области инженерии знаний и искусственного интеллекта, онтология описывается как «формальная эксплицитная спецификация совместно используемой концептуализации». Это расширение Студера и др. [31] оригинальной формулировки Грубера [18] является определением, используемым в настоящем стандарте. В контексте моделирования данных «онтология» часто интерпретируется как использование формального языка для определения формализованного представления области знания. Среди прочих задач это позволяет проверять по онтологии согласованность утверждений (фактов) и, возможно, выводить новые знания. Онтология и набор фактов (утверждений об индивидуумах) в совокупности образуют базу знаний.

Одной из фундаментальных целей онтологии является обеспечение логического вывода, включая следующие общие задачи:

- вывод принадлежности индивидуумов классам;
- вывод отношений между классами и свойствами;
- проверка согласованности базы знаний.

Пример — В области медицины онтология заболеваний может:

- из наблюдений симптомов пациента вывести характер заболевания;

- обнаружить, что различные симптомы, представленные двумя пациентами, были вызваны одним и тем же вирусом.

21.1.2 Место и роль в информационном поиске

В то время как роль большинства словарей, описанных в этом стандарте, состоит в организации выбора поисковых и индексных терминов или просмотра упорядоченных коллекций документов, онтологии в контексте поиска имеют другое назначение. Онтологии предназначены не для поиска информации с помощью индексных терминов или классификационных кодов, а для создания утверждений об индивидуумах, например о реальных людях или абстрактных вещах, таких как различные процессы. Хотя внутренняя навигация обычно не является главной задачей, онтологии могут быть полезны в некоторых случаях поиска, описанных в 21.4.

21.1.3 Создание и развитие

Термин «онтология» стал популярным благодаря его применению в области компьютерных наук (в частности, искусственного интеллекта и инженерии знаний) с начала 1990-х годов для обозначения формального описания совокупности знаний, которое можно использовать для логических рассуждений. С тех пор много исследований было посвящено разработке формальных языков для описания онтологий и обеспечения автоматизации логических рассуждений (см. также 21.1.6).

В области биомедицины некоторые сложные онтологии интегрируются с другими онтологиями в той же или смежных областях и обеспечивают результаты логических рассуждений, которые имеют правильное и полезное соответствие реальной жизни [29]. В этой области онтология, больше, чем любой другой тип словаря, предполагает тщательный анализ природы реальных сущностей, представленных в ней. Таким образом, онтологии как артефакты компьютерных наук отчасти вбирают в себя первоначальный смысл термина «онтология», т. е. раздел философии, также известный как общая метафизика или наука о бытии.

Позже в литературе по Семантическому вебу стал использоваться термин «облегченная онтология» для охвата всех видов структурированных словарей и систем организации знаний, включая тезаурусы, классификационные системы и т. п. Этот термин не используется в настоящем стандарте, поскольку в результате свободного употребления данного термина размываются границы понятий, что нецелесообразно. Руководство по использованию *Простые системы организации знаний* (Simple Knowledge Organization Systems, SKOS [37]) *для публикации и подключения тезаурусов (и других систем организации знаний) к Семантическому вебу см. в ГОСТ Р 7.0.91—2015, раздел 17.*

21.1.4 Словарный контроль

Словарный контроль (как определено в 3.98) не является собственно целью онтологий. Например, они не всегда заботятся об устранении неоднозначности омографов, чего следует ожидать в тезаурусе. В онтологиях устранение неоднозначности терминов менее важно, поскольку и классы, и индивидуумы однозначно идентифицируются с помощью средств, отличных от естественного языка. Тем не менее полезность однозначной маркировки часто признается в научной литературе по онтологиям.

21.1.5 Типы онтологий

Онтологии могут отличаться, например, степенью специфичности, областью применения или целью применения [19]. Так, онтологии наивысшего уровня определяют наиболее общие, независимые от области знания категории бытия, тогда как онтологии отдельных областей и задач описывают, соответственно, классы в конкретной области и классы для конкретной задачи, иногда основанные на онтологиях наивысшего уровня.

Хотя эти различия не являются четкими, настоящий стандарт ограничивается рассмотрением онтологий формальной области, которые имеют сложность и охват, сопоставимые со структурированными словарями, обычно используемыми для поиска информации.

21.1.6 Логика и языки представления онтологий

Как правило, логика предикатов первого порядка (или ее подмножество) используется для выражения элементов онтологий. Для описания онтологий доступны различные языки представления. Примерами являются «Система описания ресурсов (RDFS)» [36] или «Сетевой язык онтологий (OWL)» [35], которые рекомендованы Консорциумом WWW (W3C). OWL предоставляет набор аксиом, который преднамеренно ограничен, чтобы можно было использовать приемлемые алгоритмы логических рассуждений.

21.2 Семантические компоненты и отношения

21.2.1 Обзор

Для обсуждения компонентов онтологий принимается терминология OWL. Основными компонентами являются классы, свойства, аксиомы и индивиды (объекты). Некоторые другие особенности онтологий также кратко описаны в этом разделе.

Примечание — В иллюстративном примере в 21.2.9 используется соглашение об именовании CamelCase. Код UpperCamelCase используется для обозначения идентификаторов классов, а LowerCamelCase — для идентификаторов отношений. Для удобства чтения могут быть дополнительно предусмотрены метки на естественном языке в соответствии с особым соглашением о маркировке.

21.2.2 Классы

В онтологии класс — это конструкт с набором ограничений на свойства индивидов, которые образуют критерии членства в классе. В отличие от систем классификации, онтологии используют формальный язык для выражения свойств, которые служат для эксплицитного определения классов. Так, онтология может определять классы по отношению к другим классам, используя логические связки и другие ограничения. Например, класс может быть определен как пересечение двух других классов (обозначение индивидов, принадлежащих сразу к обоим классам) или как дополнение к другому классу (обозначение индивидов, не принадлежащих к этому классу).

21.2.3 Свойства

Каждый класс может быть описан свойствами его членов (индивидов). Свойствами могут быть:

- атрибуты (например, «имеетНазвание», «имеетМассу»), которым могут быть назначены конкретные значения (например, «50 грамм»);
- отношения между членами одного класса и членами других классов (например, отношение «планетаВращаетсяВокруг» в утверждении «Земля планетаВращаетсяВокруг Солнце»; см. пример в 21.2.9).

Использование свойств может быть ограничено. Например, для класса «Планета» в примере в 21.2.9 может быть указано, что атрибут «имеетМассу» имеет только одно значение.

21.2.4 Аксиомы

Аксиомы — это утверждения, определяющие основные качества классов, свойств и других сущностей в онтологии. Аксиома может быть такой же простой, как утверждение, что существует определенный класс или свойство. Другие примеры аксиоматических утверждений описывают свойства класса (например, класс «Планета» должен иметь атрибут «имеетМассу»).

21.2.5 Иерархии классов

Иерархическая структура в онтологиях задается посредством аксиом подчинения подклассов надклассам. Подклассовое отношение подразумевает, что все аксиоматические утверждения родительского класса (надкласса) также применяются к его дочерним классам (подклассам), так что ограничения свойств наследуются всеми дочерними классами. Подклассовое отношение является транзитивным, то есть распространяется вниз на подклассы всех уровней.

21.2.6 Индивиды

Индивиды — это объекты рассмотрения в определенной области знания, о свойствах которых онтология делает определенные утверждения. Индивиды также называются экземплярами и реализациями классов. Примерами индивидов являются конкретный человек, конкретная печатная книга, а также абстрактные вещи, такие как определенный симптом пациента, определенный процесс или событие, или индивидуальное возникновение чувства, такого как любовь.

21.2.7 Утверждения

Утверждения — это особая группа аксиом, которые являются сообщениями об отдельных индивидах в домене. В частности, «утверждение о классе» гласит, что индивид является членом некоторого класса. Другие утверждения могут утверждать, что два индивида одинаковы или имеют определенные ограничения на свойства. Утверждения и индивиды обычно вводятся в приложения онтологии, а не в ее структуру. Тем не менее и те, и другие могут быть описаны языком OWL, а утверждения должны соответствовать другим аксиомам онтологии.

21.2.8 Метки и идентификаторы

Классы, свойства и индивиды имеют идентификаторы. Хотя метки на естественном языке не являются строго необходимыми в онтологиях, они также часто предоставляются для улучшения читаемости.

21.2.9 Иллюстрация простой онтологии

В таблицах 13—15 показаны некоторые ключевые особенности онтологий. Графическая иллюстрация показана на рисунке 17.

Примечание — Этот очень упрощенный пример, который имеет целью показать некоторые ключевые особенности онтологий по сравнению с тезаурусами.

Таблица 13 — Классы фрагмента онтологии в области астрономии

Имя класса	Свойства/Аксиомы	
АстрономическийОбъект	имеетМассу:	тип данных
Планета	подклассКласса:	АстрономическийОбъект
	планетаВращаетсяВокруг:	*некая* Звезда
Звезда	подклассКласса:	АстрономическийОбъект

Таблица 14 — Некоторые свойства фрагмента онтологии в области астрономии

Имя свойства	Аксиомы	
имеетМассу	Область определения:	АстрономическийОбъект
	Область значений:	тип данных Масса (в кг)
вращаетсяВокруг	Область определения:	АстрономическийОбъект
	Область значений:	АстрономическийОбъект
	Инверсное отношение:	имеетНаОрбите

Окончание таблицы 14

Имя свойства	Аксиомы	
планетаВращаетсяВокруг	Область определения:	Планета
	Область значений:	Звезда
	Инверсное отношение:	имеетНаОрбитеПланету
имеетНаОрбите	Область определения:	Астрономический объект
	Область значений:	Астрономический объект
	Инверсное отношение:	вращаетсяВокруг
имеетНаОрбитеПланету	Область определения:	Звезда
	Область значений:	Планета
	Инверсное отношение:	планетаВращаетсяВокруг

Таблица 15 — Примеры индивидов (экземпляров) в области астрономии

Имя объекта	Утверждения	
Земля	экземплярКласса:	Планета
	планетаВращаетсяВокруг:	Солнце
	имеетМассу:	$5{,}972 \cdot 10^{24}$ кг
Солнце	экземплярКласса:	Звезда
	имеетНаОрбитеПланету:	Земля
	имеетМассу:	$1{,}989 \cdot 10^{30}$ кг

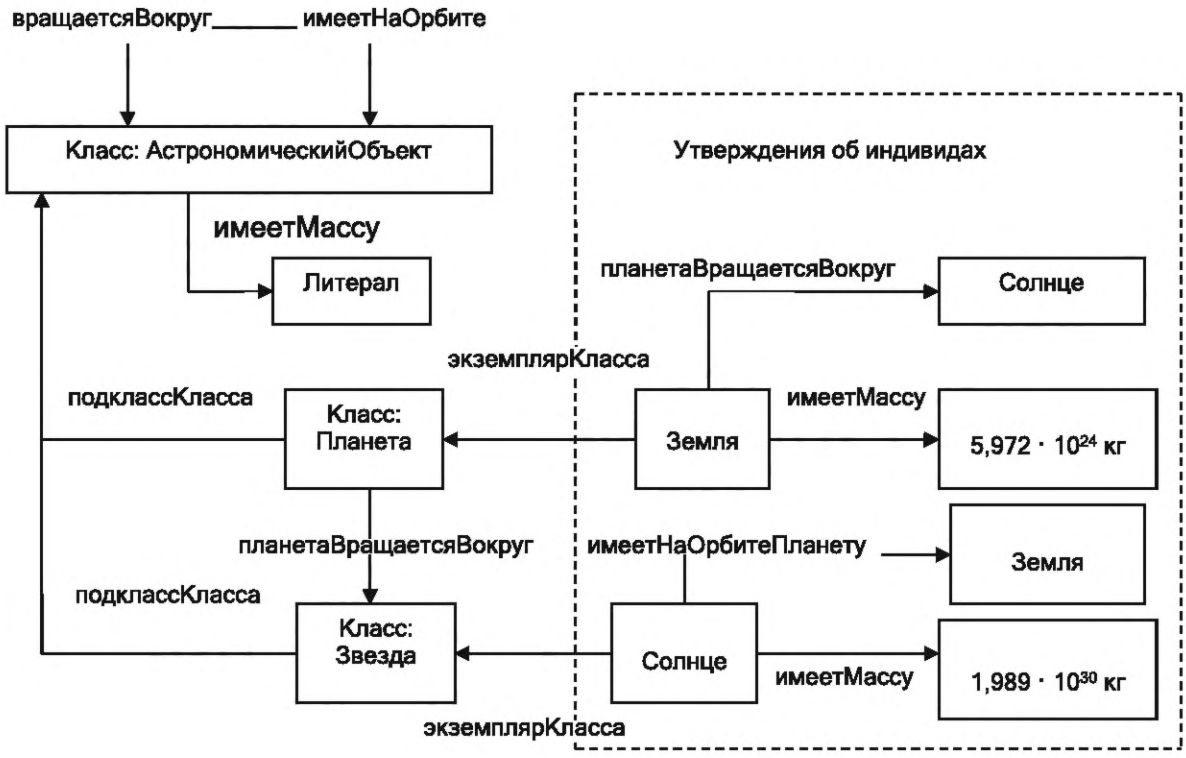

Рисунок 17 — Иллюстрация фрагмента онтологии и утверждений

В этом примере все подклассы «АстрономическийОбъект», такие как «Планета» или «Звезда», наследуют все свойства своего суперкласса. Эта онтология утверждает, что каждый астрономический объект должен иметь свойство массы, и некоторые астрономические объекты могут вращаться вокруг другого астрономического объекта. Поскольку «Планета» является подклассом класса «АстрономическийОбъект», можно сделать вывод, что планета имеет массу и может вращаться вокруг другого астрономического объекта. Кроме того, онтология допускает ограничения на использование свойств. Здесь, вводя подсвойство «планетаВращаетсяВокруг», онтология требует, чтобы все планеты вращались вокруг звезд.

21.3 Структурное сравнение тезаурусов и онтологий

Одно из ключевых отличий онтологий от тезаурусов состоит в том, что для обеспечения возможности логических рассуждений и умозаключений онтологии обязательно различают классы и индивиды.

Пример — «Земля» и «планеты» могут быть двумя понятиями в тезаурусе, связанными иерархическими отношениями выше/ниже. Инстанциональный характер взаимосвязи можно показать, используя отношения выше-множество/ниже-элемент (метки BTI и NTI в английском языке и метки «вм» и «нэ» в русском языке). А в онтологии по небесной механике «Планета» может быть определена как класс, а «Земля» — как индивид (поскольку последняя уникальна в своем домене, тогда как планет существует более одной). Эти термины могут быть связаны утверждением о принадлежности классу. Наличие таких утверждений и аксиом позволяет индивиду «наследовать» все свойства классов, к которым он принадлежит.

Понятия в тезаурусе и классы в онтологии представляют значения двумя принципиально разными способами. Тезаурусы выражают значения понятий через термины, опираясь на дополнительные данные — иерархию, связанные понятия, реляторы, лексические примечания и/или точные определения — и все эти данные предназначены в основном для пользователей. Онтологии, напротив, передают значение классов через машиночитаемые элементы.

В тезаурусе каждое понятие имеет предпочтительный термин для его представления (см. ГОСТ Р 7.0.91—2015, подраздел 4.1) и часто имеет один или несколько непредпочтительных терминов для поддержки полноты поиска. Соответственно, в онтологии каждый класс и индивид обычно (хотя и не всегда) имеет метку. Классу можно назначать несколько меток, и необходимости различать предпочтительный и непредпочтительный статус терминов обычно не бывает, поскольку предоставление контролируемых точек доступа для индексирования и поиска не является основной целью онтологий.

Устранение неоднозначности терминов естественного языка является более важной проблемой для тезаурусов, чем для онтологий. Например, потенциальный термин «луна» может иметь более одного значения. В тезаурусе этой неоднозначности следует избегать, например, используя множественное число «луны» или «естественные спутники» для применения к лунам любой планеты, и сохраняя единственное число собственного имени «Луна» для конкретного спутника, который вращается вокруг Земли. Если бы оба эти понятия были включены в одну онтологию, у них были бы разные идентификаторы, и поэтому различие форм терминов не было бы существенно.

Как видно в приведенном выше примере, отношение экземплярности, используемое в некоторых тезаурусах, приближается по смыслу к утверждениям о членстве в классе, используемым в онтологиях. Аналогично, родо-видовое иерархическое отношение (выше-род/ниже-вид, которое можно установить в тезаурусе между терминами «планеты» и «небесные тела») соответствует аксиоме подкласса в онтологиях. Однако на практике лишь немногие тезаурусы проводят различие между родо-видовыми отношениями, отношениями целое-часть и множество-элемент (***см. ГОСТ Р 7.0.91—2015, подраздел 10.2***). Недифференцированные иерархические отношения, наиболее часто встречающиеся в тезаурусах, недостаточны для функций логических рассуждений в онтологиях. *Точно так же ассоциативные отношения не подходят для онтологии, потому что они используются во множестве различных ситуаций (см. ГОСТ Р 7.0.91—2015, пункт 10.3.3) и, следовательно, не являются семантически точными для того, чтобы делать логические выводы.*

21.4 Совместимость с тезаурусами

21.4.1 Способы применения

В других разделах настоящего стандарта основной рассматриваемой формой взаимодействия является сопоставление с целью обеспечения возможности преобразования поисковых и/или индексных

терминов. В случае онтологий, однако, приложения для поиска информации, использующие онтологии непосредственно для индексирования, являются нетипичными, и поэтому сопоставление в целях индексирования здесь не рассматривается. Вместо этого в этом разделе рассматриваются следующие три варианта применения, которые входят в число используемых в настоящее время возможностей:

a) преобразование тезауруса в онтологию;

b) совместное использование тезауруса и онтологии;

c) гибридные подходы.

21.4.2 Преобразование тезауруса в онтологию

Целью преобразования тезауруса в онтологию является создание онтологии, которая включала бы в себя знания, заложенные в тезаурус, и использовала бы их в логических операциях. В некоторых случаях может быть поставлена цель объединить тезаурус с существующей формальной онтологией верхнего уровня, так чтобы классы в объединенном продукте извлекались из исходной онтологии на верхних уровнях и из исходного тезауруса на нижних уровнях.

Чтобы поддержать преобразование тезауруса в онтологию, следует выбрать подходящий формальный язык и другие инструменты онтологии. Классы и индивиды, необходимые для онтологии, должны быть выбраны среди понятий тезауруса и (возможно) дополнены новыми классами и индивидами. Это скорее всего приведет к изменению иерархической структуры тезауруса. Понятия и отношения в тезаурусе должны быть изучены и переформулированы в мельчайших деталях, где это необходимо, чтобы устранить какую-либо двусмысленность, наносящую ущерб логическим выводам онтологии. Может потребоваться различать виды иерархических отношений, если в тезаурусе присутствует только обобщенное отношение выше/ниже. Дополнительные аксиомы и другие особенности онтологии должны быть добавлены по мере необходимости.

Получившаяся в результате онтология, сопровождаемая набором утверждений, могла бы обеспечить ряд новых приложений возможностями логического вывода. Однако некоторые качества исходного тезауруса, предназначенного для поддержки индексаторов и пользователей, осуществляющих поиск, могут быть потеряны в процессе преобразования тезауруса. Нельзя гарантировать эффективность использования полученного продукта для информационного поиска.

Прежде чем приступить к проекту преобразования, следует рассмотреть возможность применения новых приложений, связанных с логическими рассуждениями. Должны быть рассмотрены вероятные затраты/выгоды; принесут ли затраты времени и ресурсов достаточную выгоду в новых приложениях или улучшение производительности? В некоторых случаях может оказаться возможным использовать существующий тезаурус в сочетании с существующей онтологией, как описано в приведенных ниже примерах. Стоимость переработки большого тезауруса не следует недооценивать.

21.4.3 Совместное использование тезауруса и онтологии

Интерес к концепции «связанных данных» привел к работам по многим проектам (см., например, Итоговый отчет Группы развития библиотечных связанных данных Консорциума Всемирной паутины [9]), которые объединяют логические способности онтологий с поисковыми способностями систем организации знаний, таких как тезаурусы. (Руководство и варианты использования для сообщества пользователей библиотек приведены в Справочном руководстве [9].) В настоящее время имеются примеры использования следующих двух сценариев, и в обоих из них широко применяются стандарты Семантического веба, такие как SKOS [37] и OWL [35]:

- Тезаурус может использоваться для индексирования элементов базы знаний, чтобы поддерживать поиск и/или просмотр базы знаний и ее онтологии. Значимая справочная информация для классов онтологии также может быть взята из сопроводительного набора документов, проиндексированных по тезаурусу. Этот подход ориентирован на пользователя, который опирается на естественный язык, а не на формализованный искусственный язык онтологии;

- Совместное использование тезауруса и онтологии может быть обеспечено набором элементов метаданных (также известным как система метаданных). При таком подходе должны быть изучены определения и атрибуты элементов метаданных и разработана модель предметной области. Следует разработать онтологию, в которой выбранные элементы метаданных устанавливаются как классы и/или свойства. Область значений элементов метаданных определяется используемым контролируемым словарем, например тезаурусом. Эта комбинация потенциально позволяет делать логические выводы на уровне элементов метаданных, а поиск информации — проводить через термины и понятия контролируемых словарей.

Пример — Модель данных Europeana Data Model (EDM) [20] основана на онтологиях и системах метаданных, таких как ORE [25], FOAF [9] и Dublin Core [17]. Она рекомендует применять общедоступные тезаурусы и другие нормативные списки для значений, которые будут использоваться элементами метаданных Subject (Тема), Creator (Создатель) и др.

21.4.4 Гибридные подходы

Имеется возможность использовать выбранные понятия и отношения из существующего тезауруса для получения новой онтологии, не отбрасывая первоначальный тезаурус. В этом сценарии элементы из одной или нескольких схем метаданных также могут быть интегрированы в онтологию. Такой проект включает в себя как преобразование тезауруса в онтологию, так и использование его совместно с получившейся в результате онтологией.

Появление Семантического веба и связанных с ним инструментов онтологии стимулировало всплеск научно-исследовательских и опытно-конструкторских разработок. Следовательно, этот список вариантов использования является открытым, и можно ожидать появления новых потребностей и возможностей для взаимодействия тезаурусов и онтологий.

22 Терминосистемы

22.1 Ключевые характеристики и происхождение

22.1.1 Общее описание

Терминосистема — это «набор обозначений, принадлежащих одному специальному языку» (см. [3], пункт 3.5.1), причем каждое обозначение представляет понятие посредством знака, символа, термина или наименования.

Терминологические данные могут быть представлены в различных форматах терминосистем, таких как банки терминов, терминологические базы, глоссарии и др.

22.1.2 Место и роль в информационном поиске

Традиционная цель терминосистем, таких как глоссарии и терминологические базы, состоит в том, чтобы улучшить общение в устной или письменной речи. Например, терминосистема может помочь переводчику или автору найти точный термин, понять значение термина или найти разные термины для конкретного понятия на одном или на другом языке. Поэтому, в отличие от тезаурусов и большинства других типов словарей, описанных в настоящем стандарте, поиск информации не является основной мотивацией создания терминосистем. Тем не менее термины из терминосистемы могут быть полезны при поиске по полному тексту, может быть, с использованием инструментов обработки естественного языка, которые не входят в сферу рассмотрения настоящего стандарта.

22.1.3 Создание и развитие

История терминологической работы уходит корнями в исследование таксономистов семнадцатого и восемнадцатого веков, которые стремились последовательно и однозначно характеризовать и называть такие объекты, как растения и животные. Из этой работы выросло видение научного языка как отличного от общего языка. В последнее время стандартизация терминологии согласно стандартам ИСО/ТК 37 направлена на обеспечение разработки качественных информационных ресурсов и облегчение их использования. В [1] приведены принципы терминологической работы. В [6], часто называемом TermBase eXchange (TBX), описаны системы управления терминологическими данными и определен формат представления структурированных терминологических данных, что в частности актуально для данного стандарта.

В XXI веке становится все более распространенным создание предприятиями терминосистем для всего спектра деловых операций с целью улучшения согласованности и качества обмена информацией и ее публикации.

22.1.4 Словарный контроль

Поскольку терминосистемы обычно не предназначены для индексирования, отсутствует необходимость использовать их в словарном контроле, как он определен в этом стандарте. Тем не менее, важной целью терминосистем является поддержка последовательности в использовании терминов и, как следствие, предоставление их определений и эквивалентов на других языках.

22.1.5 Типы терминосистем

К терминосистемам относятся совокупности ресурсов с разной степенью контроля и нормативности, от неконтролируемого словаря до стандартизированных терминосистем, созданных как словарные ресурсы на основе консенсуса и используемых, например, в документации различного рода.

22.2 Семантические компоненты и отношения в сравнении с аналогичными компонентами и отношениями тезауруса

22.2.1 Обзор

В 22.2.2—22.2.5 акцентировано внимание на некоторых ключевых сходствах и различиях между терминосистемами и тезаурусами. Дополнительно терминосистемы могут включать примечания, тематические пометы, языковые идентификаторы, идентификаторы источников, контекст и многие другие компоненты, которые в стандарте подробно не описаны.

22.2.2 Понятия

Терминологи считают «понятие» единицей знаний и единицей мышления. В целом этот взгляд совместим с определением, приведенным в настоящем стандарте. Однако ГОСТ Р ИСО 704 различает два типа понятий:

- «единичное понятие» соответствует только одному объекту и в целом сопоставимо с типом сущностей, которым могут быть даны имена в нормативном списке имен (см. раздел 23);
- «общее понятие» соответствует двум или более «объектам» (включая как абстрактные понятия, так и физические объекты).

Большинство понятий в тезаурусе будут считаться «общими понятиями» в терминосистеме, хотя тезаурус может также допускать и «единичные понятия».

22.2.3 Определения

Для терминолога определение представляет собой способ представления понятия, который позволяет отличить это понятие от других в рамках предметной области. *Определения обычно в тезаурусы не включают, но когда они включены (как показано в модели данных в ГОСТ Р 7.0.91—2015, раздел 15), они соответствуют примененным терминам, а не понятиям.* Таким образом, функционально «определение» в терминосистеме более точно соответствует лексическому примечанию тезауруса, цель которого — кратко прояснить семантические границы понятия.

22.2.4 Обозначения

В терминосистеме «понятие» представлено «обозначениями», которые соответствуют терминам тезауруса. Однако ***ГОСТ Р ИСО 704*** признает «терминами» только обозначения «общих понятий». Обозначения единичных понятий известны как наименования, сравнимые с «именами», описанными в разделе 23. Например, «памятник» может быть термином, представляющим общее понятие, в то время как «Эйфелева башня» является наименованием, представляющим единичное понятие.

Когда возникают случаи синонимии, ***ГОСТ Р ИСО 704*** рекомендует, чтобы один из синонимов был установлен как «предпочтительный термин», в то время как приемлемые синонимы являются просто «допустимыми терминами», а неприемлемые синонимы становятся «нерекомендованными терминами». В тезаурусе последние два могут рассматриваться как непредпочтительные термины (аскрипторы) с добавлением релятора при необходимости.

Тезаурусная практика добавления реляторов в скобках (***см. ГОСТ Р 7.0.91—2015, пункт 6.2.2***) для разрешения неоднозначности терминов не является обычной в терминосистемах.

Терминологическая работа может включать оценку «рейтинга приемлемости термина» по шкале, отражающей степень, в которой термин точно обозначает понятие. В ходе сопоставления терминосистемы и тезауруса этот рейтинг может быть полезен при определении необходимости маркера точной или неточной эквивалентности (см. раздел 11).

22.2.5 Отношения

ГОСТ Р ИСО 704 рекомендует следующие «отношения», сопоставимые с отношениями тезауруса:

- **иерархические отношения** между двумя понятиями. ***ГОСТ Р ИСО 704*** признает два типа иерархических отношений — родо-видовое и партитивное. В целом они близко соответствуют родо-видовым (род—вид) и партитивным (целое—часть) иерархическим отношениям тезауруса соответственно (***см. ГОСТ Р 7.0.91—2015, подраздел 10.2***). Иерархические инстанциональные отношения (множество—элемент), определенные в ***ГОСТ Р 7.0.91***, не имеют параллелей в ***ГОСТ Р ИСО 704*** (поскольку экземпляры объектов, отражаемых понятием, трактуются в терминосистемах иначе);
- **ассоциативные отношения** между понятиями почти такие же, как ассоциативные отношения тезауруса, и так же различаются по качеству (***см. ГОСТ Р 7.0.91—2015, подраздел 10.3***);
- **синонимия** и **квазисинонимия** между терминами на одном языке приблизительно равны «отношению эквивалентности» на одном естественном языке тезауруса (***см. ГОСТ Р 7.0.91—2015, раздел 8***). В терминологической работе, напротив, «эквивалентность» обычно применяется только в контексте терминов, взятых из разных языков.

Кроме того, ***ГОСТ Р ИСО 704—2010 (раздел 5)*** подчеркивает необходимость отнесения понятия к определенной предметной области, в которой оно определено. Это имеет параллель в работе с тезаурусом, где понятия часто устанавливают в соответствии с потребностями конкретного применения, хотя это и не выражается формальными средствами. *Как изложено в разделе 5 ГОСТ Р 7.0.91—2015, содержание понятия не всегда соответствует самому общему значению связанного с ним предпочтительного термина.*

22.3 Совместимость с тезаурусами

22.3.1 Общие положения

Компании, стремящиеся к гармоничному управлению всеми своими терминологическими и словарными ресурсами, иногда создают корпоративное хранилище для совместного использования информационных ресурсов. Некоторые из случаев взаимодействия описаны в 22.3.2—22.3.4.

22.3.2 Сопоставление с тезаурусом

Поскольку терминосистемы не используются для индексирования коллекций ресурсов, сопоставление терминосистемы с тезаурусом не имеет таких целей, как для большинства словарей, описанных в настоящем стандарте. Соответствия не предназначены для прямого преобразования индексных терминов или запросов для поиска по метаданным. Однако существуют альтернативные варианты использования, которые требуют селективного соответствия (см. 6.5) общих и/или единичных понятий терминосистемы соответствующим дескрипторам тезауруса.

На рабочем уровне установление соответствия понятий между терминосистемой и тезаурусом должно следовать той же общей методологии и практике, что и между двумя тезаурусами (см. разделы 7—12).

Особое внимание следует уделить сфере действия системы понятий, охватываемой терминосистемой.

Во всех утверждениях о соответствии понятие должно быть однозначно обозначено с использованием предпочтительного термина (наименования) или несемантического уникального идентификатора. Первый больше подходит для читателей, второй — для операций на компьютере.

22.3.3 Расширение тезауруса в области, охватываемой терминосистемой

Терминосистемы могут быть полезны в качестве источника понятий и/или терминов при создании или поддержке тезауруса. Они также могут помочь при написании лексических примечаний, при выборе предпочтительного термина и при создании определений терминов, если это необходимо (***подробное руководство см. ГОСТ Р 7.0.91—2015, разделы 5, 6 и пункт 13.2.2***). Важно помнить, что записи в терминосистеме, как правило, выбираются с учетом потребностей определенной предметной области. Когда термины применяются в другом контексте, могут потребоваться корректировки.

Понятия и термины, выбранные для включения в тезаурус, должны тщательно проверяться, и к любым терминам с несколькими значениями должны добавляться реляторы в соответствии с ГОСТ Р 7.0.91—2015, пункт 6.2.2.

22.3.4 Дополнение тезауруса в поисковых приложениях

Если документы, подлежащие поиску, не были классифицированы или проиндексированы, или, если существующие метаданные не подходят для поисковых целей, может быть полезен «поисковый тезаурус». Это словарь, который не использовался для индексирования и ***может не соответствовать ГОСТ Р 7.0.91***, но способен поддерживать полнотекстовый поиск. Поиск может быть организован с использованием комбинации тезауруса и терминосистемы, и это легче осуществить, если соответствующие понятия уже были сопоставлены, как описано в 22.3.1. Для любого понятия могут использоваться все соответствующие термины и наименования терминосистемы, а также соответствующие термины тезауруса. Поисковый тезаурус иногда может принимать форму словаря синонимических рядов (см. раздел 24, особенно 24.3.3).

23 Нормативные файлы имен

23.1 Ключевые характеристики и происхождение

23.1.1 Общее описание

Нормативные списки имен, также известные как авторитетные файлы, представляют собой наборы имен сущностей. Основные цели нормативных файлов имен — уникальная идентификация имено-

ванной сущности и предоставление доступа к записи об именованной сущности по вариантным именам и формам имен. Таким образом, нормативные списки имен обычно являются контролируемыми словарями, а также могут быть структурированными словарями.

Пример — Запись из нормативного списка британских политиков.

В этом примере показана одна запись в нормативном списке имен, в которой указаны разные имена и псевдонимы, под которыми иногда известен один конкретный человек. Чтобы отличить этого человека от других с таким же именем (например, внук Уинстона Черчилля, которого также звали Уинстон Черчилль и который также является британским политиком), в записи указана дата рождения и смерти.

Черчилль, Уинстон, сэр, 1874—1965
Альтернативные имена:
Сэр Уинстон Черчилль
Спенсер-Черчилль, Уинстон Леонард
Уинстон Леонард Спенсер Черчилль
Полковник Уорден

Часто имена являются собственными именами, обозначающими отдельные объекты, и поэтому могут рассматриваться как «одноэлементные классы». Именованные сущности включают в числе прочих следующие типы:

- лица;
- организации;
- места;
- произведения искусства (включая литературные произведения и другие документы, именем которых служит заглавие).

Именованные сущности всегда являются уникальными объектами. Напротив, термины тезауруса представляют собой гораздо более широкий спектр понятий, некоторые из которых являются весьма общими и абстрактными, как описано в ***ГОСТ Р 7.0.91—2015, подраздел 5.1***.

Примеры в этом разделе будут сосредоточены на нормативных списках имен для наиболее распространенных типов сущностей, перечисленных выше, но иллюстрируемые ими принципы могут применяться более широко для других типов сущностей.

Примечание — Использование прилагательных «авторитетный», «нормативный» не подразумевает, что список имен обязательно выпускается каким-либо уполномоченным органом; скорее, это относится к использованию списка для проверки ссылок на соответствующие сущности.

23.1.2 Место и роль в информационном поиске

В контексте данного стандарта нормативный список имен может использоваться для:

- обеспечения согласованности записей в каталогах, указателях и базах данных;
- просмотра и выбора поисковых терминов при поиске информации;
- автоматического извлечения имен сущностей из текста, иногда требующего обработки метаданных и логических выводов.

Нормативные списки имен часто поддерживают контролируемый доступ к элементам метаданных документов, таким как авторы, издатели и места публикации, тематика. В этом отношении нормативные списки имен и тезаурусы выполняют одну и ту же функцию — уникально идентифицируют имена, чтобы их последовательно использовали все стороны при индексировании и поиске информации, относящейся к конкретной сущности, с целью нахождения всех релевантных документов в данной коллекции. Кроме того, структурированные нормативные списки имен могут помочь в расширении поискового запроса путем включения подчиненных объектов — юридических лиц, географических областей и т. п.

23.1.3 Создание и развитие

Традиция ведения нормативных списков имен за многие десятилетия распространилась от отдельных организаций, таких как библиотеки, до ресурсов общего пользования, таких как списки общественных организаций, общедоступные базы данных и национальные проекты. В последнее время совместные международные проекты, такие как Виртуальный международный авторитетный файл (Virtual International Authority File, VIAF) [33], поощряют владельцев отдельных списков объединять свои данные в один большой общий ресурс.

23.1.4 Типы нормативных списков имен

Нормативные списки имен варьируют от простых до сложных. Некоторые из них представляют собой простые списки, обычно обогащенные вариантами имен, в то время как другие представляют

собой структурированные словари, показывающие тезаурусоподобные иерархические, а иногда и ассоциативные отношения. Простые списки часто содержат только лексические значения контрольных записей, в то время как структурированные словари включают вложенные иерархии записей и/или автономные записи, связанные отношениями.

Как показано на примере в таблице 16, простой список представляет имена в алфавитном порядке, в то время как структурированный словарь использует отступ для отражения иерархических отношений целое—часть.

Таблица 16 — Простой и структурированный нормативные списки имен

Простой нормативный список имен	Структурированный нормативный список имен
Англия	Соединенное Королевство
Великобритания	Великобритания
Соединенное Королевство	Англия
Шотландия	Шотландия

23.1.5 Словарный контроль

Поскольку целью ведения нормативного списка является согласованное именование, среди всех имен, данных для конкретной сущности, часто выделяют предпочтительное имя (см. 23.2.2). Иногда для каждого языка указывается свое предпочтительное имя. Чтобы различать разные сущности с одинаковыми именами, некоторые списки добавляют однозначный релятор к каждому из неоднозначных имен. Однако чаще каждой сущности присваивают отдельный уникальный идентификатор и предоставляют дополнительную информацию, помогающую пользователю различать сущности (см. 23.2.3).

23.2 Семантические компоненты и отношения

23.2.1 Обзор

Основными семантическими компонентами в нормативном списке имен являются имена, дополнительная информация об именованных сущностях и отношения между именами и между сущностями. Именам и/или сущностям также может быть присвоен уникальный идентификатор, но этот компонент не является семантическим и далее рассматриваться не будет.

23.2.2 Имена

Подобно тезаурусу, нормативный список имен обычно предоставляет несколько точек доступа для поиска по всем известным наименованиям данной сущности, включая аббревиатуры, псевдонимы, различные переводные, транслитерированные, инвертированные формы; формы, определенные применением других правил, а также другие формы, полезные для поиска.

В случае организаций, для которых может быть представлена иерархия подразделений, иногда имя родительской организации может быть объединено с именем подразделения, чтобы сформировать уникальное однозначное имя для каждого подразделения. Например, генеральный директорат Европейской комиссии «Информационное общество и СМИ» может быть представлен в нормативном списке имен как «Европейская комиссия. Информационное общество и СМИ».

23.2.3 Дополнительная информация о сущности

Каждая запись в нормативном списке имен обычно имеет дополнительную информацию, связанную с сущностью и ее именами. Эта информация служит для различения сущностей с одинаковыми именами. Отличительная информация может быть добавлена непосредственно к имени в качестве релятора или же добавлена к записи в качестве атрибута или примечания.

В зависимости от типа сущности дополнительная информация может включать:

- дату и место рождения/смерти или основания/ликвидации, периода деятельности или семейных отношений (в случае лиц или организаций);
- географические координаты или типы местности (в случае мест);
- аффилиацию, семейные отношения, роль или гражданство (в случае лиц или организаций);
- создателя или дату создания (в случае произведений);
- периоды, места или языки, в которых используется/использовалось данное имя;
- основные работы или виды деятельности, благодаря которым лицо стало известно.

23.2.4 Отношения

Обычно встречаются следующие типы отношений:

a) отношения эквивалентности, используемые между предпочтительным именем и альтернативными именами одной и той же сущности. На эти отношения указывают перекрестные ссылки, такие как «смотри/ссылка от», или ***метки USE/UF в английском языке и «см/с» в русском языке***. Данные отношения подходят для всех типов именованных сущностей и являются наиболее распространенными в нормативном списке имен;

b) иерархические отношения, часто используемые, когда сущность образует часть вышестоящего целого. Родо-видовые отношения обычно не возникают, потому что имена идентифицируют уникальные сущности, а не классы. Иерархические отношения могут быть отражены в самом имени или указаны уровнем отступа, типографскими способами или явными ***метками, такими как BT/NT или BTP/NTP в английском языке и «в/н» или «вц/нч» в русском языке (см. ГОСТ Р 7.0.91—2015, пункт 10.2.3)***. Иерархические отношения чаще всего встречаются в списках географических названий и наименований организаций;

c) ассоциативные отношения, которые обозначены перекрестными ссылками, такими как «смотри также», или ***меткой RT в английском языке и «а» в русском языке***. Они используются для связи сущностей разного типа, например для выражения аффилиации лица с организацией и наоборот;

d) хронологические отношения, также известные как «отношения следования», используются в тех случаях, когда идентичность сущности менялась с течением времени, как правило, сопровождаясь сменой имени. На эти отношения указывают различные выражения, такие как «впоследствии/ранее» или недифференцированная ссылка «смотри также». Эта связь чаще всего встречается в списках наименований организаций и географических названий;

e) сложные отношения, когда одна организация разделяется на две или более или когда две или более отдельных организаций объединяются в одну, обычно встречаются в списках наименований организаций. Разделение и слияние также могут рассматриваться как тип хронологических отношений, с тем осложнением, что отношением связаны более двух объектов. Эти отношения иногда помечаются перекрестными ссылками «впоследствии/ранее» и/или примечанием с качественным объяснением изменения.

23.3 Сопоставление тезауруса и нормативного списка имен

23.3.1 Общие положения

Необходимость подготовки полного соответствия всех записей тезауруса записям нормативного списка имен (или наоборот) возникает редко из-за различия целей и содержания этих словарей. Тезаурус, как правило, содержит множество понятий, которые неуместно включать в список нормативных имен. И если список имен достаточно велик, в тезаурусе может появиться только меньшинство его записей. Поэтому селективное соответствие (см. 6.5) является наиболее вероятным подходом.

Все типы соответствий, описанные в разделах 7—11, могут рассматриваться при сопоставлении тезауруса и нормативного списка имен, но обычно наиболее подходящим типом является точная эквивалентность. Это связано с тем, что собственные имена относятся к однозначно идентифицируемым индивидам, и они в таком же качестве могут встречаться в тезаурусе как дескрипторы.

Иерархическое соответствие или неточная эквивалентность будут уместны, если тезаурус и нормативный список имен определяют одну и ту же сущность с различной детализацией.

Соответствие сложной эквивалентности подходит только к случаям разделения и слияния [см. 23.2.4 d)].

Во всех утверждениях о соответствии сущность должна быть однозначно обозначена с использованием предпочтительного имени или несемантического уникального идентификатора. Первый вариант, как правило, больше подходит для читателей, последний — для операций на компьютере.

23.3.2 Практические примеры соответствия

Во всех утверждениях о соответствии в следующих примерах «ТЕЗ» является идентификатором тезауруса, а «НОРМ» — идентификатором нормативного списка имен.

Примеры:
1

Случай соответствия	*От тезауруса, содержащего понятие «Уинстон Черчилль (1874—1965)», к нормативному списку имен, указанному в 23.1.1*
Утверждение о соответствии	*Уинстон Черчилль (1874—1965) НОРМ =ЭК Уинстон Черчилль, сэр, 1874—1965* *или* *Уинстон Черчилль (1874—1965) НОРМ ЭК Уинстон Черчилль, сэр, 1874—1965*
Обсуждение	*Оба этих соответствия реверсивны (действуют в обе стороны). Хотя применение маркера точной эквивалентности «=» факультативно, его указание полезно, поскольку подкрепляет совместимость в условиях, когда взаимодействуют несколько словарей*

2

Случай соответствия	*От тезауруса, содержащего дескриптор «политики», но не содержащего нижестоящих к нему понятий, к нормативному списку имен, указанному в 23.1.1*
Утверждение о соответствии	*политики НОРМ УС Уинстон Черчилль, сэр, 1874—1965* *или* *политики НОРМ УСЭ Уинстон Черчилль, сэр, 1874—1965*
Обсуждение	*Обратные соответствия могут быть получены путем реверсии утверждений, замены идентификатора НОРМ идентификатором ТЕЗ и замены «УС/УСЭ» на «ШС/ШСМ». Хотя указывать инстанциональные отношения (УСЭ/ШСМ) не обязательно, это способствует совместимости в приложениях, где реализован логический вывод с помощью онтологий.* *Если все сущности в этом нормативном списке имен являются политиками, для каждого из них будет действительно иерархическое соответствие*

3

Случай соответствия	*От тезауруса, содержащего дескриптор «международные организации» и более узкое понятие «Организация Объединенных Наций», к нормативному списку имен, в котором есть запись «Верховный комиссар по делам беженцев», а также запись его головной организации «Организация Объединенных Наций»*
Утверждение о соответствии	*Организация Объединенных Наций НОРМ ЭК* *Организация Объединенных Наций* *или же* *Организация Объединенных Наций НОРМ =ЭК* *Организация Объединенных Наций*
Обсуждение	*Как и в примере 1, эти соответствия являются обратимыми, и рекомендуется использовать маркер «=».* *Хотя добавить иерархическое соответствие между «Организация Объединенных Наций» и «Верховный комиссар по делам беженцев» и не было бы неправильным, это обычно не требуется делать, если иерархические отношения между этими структурами показаны внутри нормативного списка имен*

4

Случай соответствия	*От тезауруса, содержащего понятие «международные организации» и более узкое понятие «Верховный комиссар по делам беженцев», к нормативному списку имен, в котором есть запись для «Организации Объединенных Наций», но не для какого-либо из ее учреждений*
Утверждения о соответствии	*Международные организации НОРМ УС Организация Объединенных Наций, а также Верховный комиссар по делам беженцев НОРМ ШС Организация Объединенных Наций или Международные организации НОРМ УСЭ Организация Объединенных Наций, а также Верховный комиссар по делам беженцев НОРМ ШСЦ Организация Объединенных Наций*
Обсуждение	*Принимая во внимание, что иерархические отношения между «международные организации» и «Организация Объединенных Наций» являются инстанциональными, отношения между «Организация Объединенных Наций» и «Верховный комиссар по делам беженцев» являются отношениями целое—часть, что обозначено метками ШСЦ/УСЧ*

5

Случай соответствия	*Соответствие понятия «Сардиния» в структурах тезауруса и нормативного списка имен:*	
—	*Структурированный нормативный список имен*	*Тезаурус*
	Европа Италия Сардиния	*острова Сардиния*
Утверждения о соответствии	*Сардиния НОРМ ЭК Сардиния или Сардиния НОРМ =ЭК Сардиния*	
Обсуждение	*Оба эти соответствия являются обратимыми. Чтобы соответствие эквивалентности между именованными сущностями было допустимым, не обязательно, чтобы иерархические структуры словарей были одинаковыми*	

6

Случай соответствия	*Соответствие для дескриптора «Ирландия» в структурах тезауруса и нормативного списка имен, показанных ниже, при этом нормативный список имен включает скорее политические образования, а не географические объекты*	
—	*Структурированный нормативный список имен*	*Тезаурус*
	Европа Ирландия	*острова Ирландия*
Утверждения о соответствии	*Ирландия НОРМ ЭК Ирландия или Ирландия НОРМ ~ЭК Ирландия*	
Обсуждение	*В этом случае уместна неточная эквивалентность, поскольку остров Ирландия включает в себя Северную Ирландию, тогда как политическое образование «Ирландия» ее не включает. Оба эти соответствия являются обратимыми*	

7

<table>
<tr><td>Случай соответствия</td><td colspan="2">Отображение между словарями, которые используют другой подход к изменению имен с течением времени. В этом случае тезаурус рассматривает римское поселение Лугдунум Батаворум как место, отличное от города Лейден, тогда как в нормативном списке имен «Лугдунум Батаворум» рассматривается как альтернативное название Лейдена</td></tr>
<tr><td rowspan="2">—</td><td>Структурированный нормативный список имен</td><td>Тезаурус</td></tr>
<tr><td>Лейден (Leiden)
Альтернативные названия
Лейден (Leiden)
Лугдунум Батаворум</td><td>Лейден (Leiden)
с Лейден (Leyden)
а Лугдунум Батаворум</td></tr>
<tr><td>Утверждение о соответствии</td><td colspan="2">Лейден ТЕЗ ЭК Лейден (Leiden) | Лугдунум Батаворум</td></tr>
<tr><td>Обсуждение</td><td colspan="2">В этом случае уместно соответствие кумулятивной сложной эквивалентности имени из нормативного списка объединению дескрипторов тезауруса. В обратном направлении применимы два иерархических соответствия:
Лейден (Leiden) НОРМ ШСЦ Лейден (Leiden);
Лейден (Leiden) НОРМ ШСЦ Лугдунум Батаворум</td></tr>
</table>

24 Словари синонимических рядов

24.1 Ключевые характеристики и происхождение

24.1.1 Общее описание

Синонимический ряд представляет понятие, перечисляя как можно больше терминов, которые можно использовать в тексте для передачи этого понятия. Поисковые системы, которые используют этот тип словаря, обычно поддерживают обширный список синонимических рядов.

Пример — Два синонимических ряда: «астронавт, космонавт, тайконавт, спатионавт, спейсмен» и «космический аппарат, космический корабль, ракета, космический шаттл».

24.1.2 Место и роль в информационном поиске

Синонимические ряды используются для поиска, но не для индексирования. Их обычно применяют к содержимому документа, а не к его метаданным. Когда поисковый запрос включает термин, присутствующий в одном из рядов, поиск расширяется, чтобы извлечь все вхождения каждого термина этого ряда. Таким образом, в примере в 24.1.1 при запросе «космонавт» будут получены ссылки на «астронавт», «тайконавт», «спатионавт», «спейсмен», равно как и на исходный термин. Применение синонимических рядов часто сочетают с алгоритмом выделения основ слов, что еще более расширяет круг найденных терминов (см. 24.2). Поскольку нет необходимости в индексировании, вновь образованный или обновленный синонимический ряд может сразу использоваться при поиске, независимо от размера или долговечности коллекции ресурсов, в которой необходимо выполнить поиск.

24.1.3 Создание и развитие

Синонимические ряды стали широко использоваться в 1990-х годах и часто используются «за кулисами» поисковых систем. Синонимические ряды включают в число контролируемых словарей (см. [7]).

24.1.4 Словарный контроль

Поскольку синонимические ряды не используются для индексирования, необходимость словарного контроля не возникает. Термины в них не имеют реляторов или других средств устранения неоднозначности, и один и тот же термин может появляться в нескольких рядах с разными значениями. Ни один из терминов не выбирают в качестве предпочтительного, и все термины в одном и том же ряду считаются эквивалентными для целей поиска.

24.2 Семантические компоненты и отношения

Основными компонентами синонимического ряда являются термины, возможно с уникальным идентификатором для каждого термина. Ряды не имеют иерархической структуры, и ни один из членов

какого-либо одного ряда не имеет «предпочтительного» статуса для поисковых целей. При желании каждому ряду может быть присвоен уникальный идентификатор.

В одном синонимическом ряду часто встречаются разные типы синонимов: истинные синонимы, устаревшие термины, сокращения, аббревиатуры, неполные синонимы, ошибочные написания и лексические варианты (***см. также ГОСТ Р 7.0.91—2015, подразделы 8.2 и 8.3***). Могут быть включены формы как множественного, так и единственного числа, хотя это и не требуется при применении алгоритма выделения основ слов.

Пример — Первый синонимический ряд примера в 24.1.1 можно расширить, включив в него: «астронавты», «космонавты», «тайконавты», «спатионавты». Когда работает алгоритм выделения основ слов, необходимо включать только неправильные формы множественного числа, такие как «человек — люди» в русском языке, «man — men» в английском языке.

Иногда в синонимический ряд включают термины более общие или более конкретные, чем другие термины ряда, хотя это может вызвать проблемы при поиске. Точность поиска также страдает, хотя и в меньшей степени, когда включены неполные синонимы и термины с несколькими значениями.

Пример — Синонимический ряд названий нового штамма вируса гриппа H1N1, вызвавшего крупную вспышку в 2009 году (широко известный как «свиной грипп»), может включать такие термины, как «свиной грипп», «вирус H1N1/09», «грипп А» и «грипп H1N1». Включение относительно широкого термина «свиной грипп» наряду с некоторыми очень конкретными терминами служит для повышения полноты поиска. Это может быть полезно в сети ресурсов, посвященных здоровью человека, но может снизить точность поиска при применении к коллекциям, содержащим документы о многих других видах свиного гриппа.

Между всеми терминами одного синонимического ряда по существу установлено отношение эквивалентности. Иерархические и ассоциативные отношения в синонимических рядах не выражаются.

24.3 Взаимодействие с тезаурусами

24.3.1 Сопоставление с тезаурусом

В период разработки этого стандарта системы для прямого сопоставления словарей синонимических рядов и тезаурусов не были распространены. Однако установление соответствий их элементов возможно, особенно если каждый ряд имеет идентификатор. При использовании идентификаторов ряда в формулировках соответствия синонимические ряды должны отображаться в соответствующие понятия тезауруса и наоборот, следуя тем же типам соответствия и рекомендациям, что и в разделах 7—11.

24.3.2 Пополнение тезауруса терминами из синонимических рядов

Синонимические ряды могут быть полезны при создании или ведении тезауруса, особенно в качестве источника дополнительных непредпочтительных терминов (аскрипторов). Подробное руководство ***см. в ГОСТ Р 7.0.91—2015, раздел 6 и пункт 13.2.2***. Термины должны быть тщательно проверены, и любые термины с несколькими значениями должны иметь реляторы, ***добавленные в соответствии с ГОСТ Р 7.0.91—2015, пункт 6.2.2***.

24.3.3 Дополнение тезауруса в поисковых приложениях

В сетевой среде пользователи часто хотят расширить поиск, используя ресурсы, которые не были проиндексированы с помощью применяемого ими тезауруса (и, возможно, не были проиндексированы или классифицированы с помощью какого-либо словаря). В этом случае необходим поиск по полному тексту, и применение словаря синонимических рядов может повысить полноту поиска. Для этого дескрипторы тезауруса следует связать с синонимическими рядами, как описано в 24.3.1.

Тот же самый метод может быть применен к ресурсам, которые были проиндексированы, но не с достаточной полнотой. Чтобы повысить полноту поиска, следует применить соответствия дескрипторов тезауруса и синонимических рядов для преобразования поисковых запросов, а преобразованные запросы использовать для поиска во всех доступных текстах.

Приложение ДА
(справочное)

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном стандарте

Таблица ДА.1

Обозначение ссылочного национального стандарта	Степень соответствия	Обозначение и наименование соответствующего международного стандарта
ГОСТ Р 7.0.91—2015 (ИСО 25964-1:2011)	MOD	ИСО 25964-1:2011 «Информация и документация. Тезаурусы и их совместимость с другими словарями. Часть 1. Тезаурусы для информационного поиска»

Примечание — В настоящей таблице использовано следующее условное обозначение степени соответствия стандартов:

- MOD — модифицированный стандарт.

Библиография

[1] ИСО 704:2022 — Терминологическая работа. Принципы и методы (ISO 704:2022, Terminology work — Principles and methods)

[2] ИСО 999:1996 — Информация и документация. Руководящие указания по содержанию, структуре и представлению указателей (ISO 999:1996, Information and documentation — Guidelines for the content, organization and presentation of indexes)

[3] ИСО 1087-1:2000 — Терминологическая работа. Словарь. Часть 1. Теория и применение (ISO 1087-1:2000, Terminology work — Vocabulary — Part 1: Theory and application)

[4] ИСО 15489-1:2016 — Информация и документация. Управление документами. Часть 1. Понятия и принципы (ISO 15489-1:2016, Information and documentation — Records management — Part 1: Concepts and principles)

[5] ISO/TR 15489-2 — Информация и документация. Управление записями. Часть 2. Руководящие указания (ISO/TR 15489-2, Information and documentation — Records management — Part 2: Guidelines)

[6] ИСО 30042:2008 — Системы управления терминологией, базами знаний и контентом — Обмен терминологическими базами [TermBase eXchange (TBX)] (ISO 30042:2008, Systems to manage terminology, knowledge and content — TermBase eXchange (TBX))

[7] ANSI/NISO Z39.19-2005 (R2010) — Рекомендации по созданию, формату и управлению одноязычными контролируемыми словарями (ANSI/NISO Z39.19-2005 (R2010), Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Available at: https://niso.org/publications/ansiniso-z3919-2005-r2010)

[8] BS 8723-4 — Структурированные словари для поиска информации. Руководство. Взаимодействие между словарями (BS 8723-4, Structured vocabularies for information retrieval. Guide. Interoperability between vocabularies)

[9] Baker T., et al. W3C Library Linked Data Incubator Group Final Report. World Wide Web Consortium, 25 October 2011. Available at: gost.e-ecolog.ru

[10] Biblioteca nazionale centrale di Firenze. Nuovo soggettario: guida al sistema italiano di indicizzazione per soggetto: prototipo del Thesaurus. Milano: Editrice Bibliografica, 2006. Availble at: gost.e-ecolog.ru

[11] Bliss H E., A bibliographic classification, extended by systematic auxiliary schedules for composite specification and notation. New York, Wilson, 1952-53. [Two previous editions, the first in 1935 published under the title: A system of bibliographic classification]

[12] Bratcher P., Smith J., Smiraglia R.P., Music subject headings: compiled from Library of Congress subject headings. Lake Crystal, MN: Soldier Creek Press, 1988

[13] Brickley D., Miller L., FOAF Vocabulary Specification, 0.98., FOAF Project, 9 August 2010. Latest version available at gost.e-ecolog.ru

[14] Brown J. D., Subject Classification: for the arrangement of libraries and the organization of information, withdables, indexes, etc., for the subdivision of subjects. 3d ed., rev. and enl., by J Douglas Stewart. London, Grafton & co., 1939. [First edition published in 1906 under the title: Subject classification, with tables, indexes, etc., for the subdivision of subjects]

[15] Cutter C.A., Expansive classification. 2nd edition. Northampton, Mass., 1902

[16] DDC 23: Dewey Decimal classification and relative index. Edition 23. Dublin, OH: OCLC, Inc., 2011. [First edition by Melvil Dewey published in 1876 under the title: A classification and subject index.] Also available as an online database, WebDewey, at: gost.e-ecolog.ru

[17] Dublin Core Metadata Initiative. Dublin core metadata element set, version 1.1. DCMI recommendation, 18 December 2006. Latest version available at gost.e-ecolog.ru [Full set of DCMI standards available at gost.e-ecolog.ru]

[18] Gruber T.R. “A translation approach to portable ontology specifications”. Knowledge acquisition. 1993, vol. 5, no. 2, pp. 199—220. Also available at: gost.e-ecolog.ru

[19] Guarino N. “Semantic matching: Formal ontological distinctions for information organization, extraction, and integration”. In: Pazienza, Maria, ed. Information Extraction A Multidisciplinary Approach to an Emerging Information Technology, International Summer School, Frascita, Italy, July 14—18, 1997. Lecture notes in computer sciences, vol. 1299. Berlin/New York: Springer: 1997, pp. 139—170

[20] Iaac A., ed. Europeana Data Model Primer, Europeana, 14 July 2013. Available at: gost.e-ecolog.ru . eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf

[21] Library of Congress Cataloging Distribution Service. LC Classification Schedules, A-Z. 41 volumes. Washington, D.C.: Library of Congress, 2008—2011. Also available as an online database at: at: gost.e-ecolog.ru

[22] Library of Congress Prints and Photographs Division. Thesaurus for graphic materials (TGM) Washington, D.C.: Library of Congress, 1995. Available at: gost.e-ecolog.ru and gost.e-ecolog.ru

[23] Library of Congress Subject Cataloging Division. Library of Congress subject headings. 33rd edition. Washington, DC.: Library of Congress, 2011. ISSN 1048-9711

[24] National Library of Medicine. Medical subject headings. Bethesda, MD: National Library of Medicine, continuously updated. Available at: gost.e-ecolog.ru

[25] ORE Specifications and User Guides. Open Archives Initiative Object Reuse and Exchange (OAI-ORE). Latest version available at: gost.e-ecolog.ru

[26] RAMEAU (Repertoire d'autorité-matière encyclopédique et alphabétique unifié). Paris: Bibliothèque nationale de France, 1980. Available at: gost.e-ecolog.ru

[27] Ranganathan S.R. Colon classification. 6th ed. New Delhi: Ess Ess Publications, 2007. [First edition published in 1933]

[28] Répertoire des vedettes-matières de l'Université Laval (RVM). Québec, QC: Bibliothèque de l'Université Laval, 1962. Available at: gost.e-ecolog.ru

[29] Gemeinsame Normdatei. Leipzig, Germany: Deutsche Bibliothek [German National Library], Available in multiple formats; see: gost.e-ecolog.ru

[30] Smith B. and Ceusters W. "Ontological realism: A methodology for coordinated evolution of scientific ontologies". Applied Ontology, Nov. 2010, vol. 5, no. 3-4, pp. 139—188

[31] Studer R., Benjamins V.R., and Fensel D., "Knowledge engineering: principles and methods". Data & knowledge engineering. 1998, vol. 25, no. 1—2, pp. 161—197

[32] Universal Decimal Classification (UDC). Complete edition. Volumes 1 & 2. BIP 0017:2006. London: British Standards Institute, August 2006. [Originally published as Manuel du Repertoire Bibliographique Universel, Brussels: IIB, 1905—1907. Currently available in over 40 languages and online. For a complete list of editions, see: gost.e-ecolog.ru]

[33] Virtual International Authority File (VIAF) [website]. Available at: gost.e-ecolog.ru

[34] Winkle L., ed. Subject Headings for Children: A List of Subject Headings Used by the Library of Congress with Dewey Numbers Added. 2nd ed. Albany, NY: Forest Press, 1998

[35] World Wide Web Consortium. OWL 2 Web Ontology Language: Document Overview (Second Edition). W3C Recommendation, 11 December 2012. Available at: gost.e-ecolog.ru

[36] World Wide Web Consortium. RDF Schema 1.1. W3C Recommendation, 25 February 2014. Available at: http:// gost.e-ecolog.ru

[37] World Wide Web Consortium. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 18 August 2009. Latest version available at gost.e-ecolog.ru

УДК 025.4:006.354 ОКС 01.140.20

Ключевые слова: Система стандартов по информации, библиотечному и издательскому делу, тезаурусы, совместимость, поиск информации, соответствия, классификационные системы, таксономии, словари предметных рубрик, онтологии, терминосистемы

Редактор *Л.В. Коретникова*
Технический редактор *В.Н. Прусакова*
Корректор *Л.С. Лысенко*
Компьютерная верстка *И.Ю. Литовкиной*

Сдано в набор 04.03.2024. Подписано в печать 21.03.2024. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 10,23. Уч-изд. л. 9,68.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «Институт стандартизации»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.
gost.e-ecolog.ru gost.e-ecolog.ru

gost.e-ecolog.ru