# Рекомендации к составлению тезауруса

1) Построение тезауруса должно идти по следующему плану:

- определение тематического охвата;

- сбор массива лексических единиц;

- формирование словника тезауруса;

- построение словарных статей и указателей;

- оформление;

- экспертиза и регистрация тезауруса.

2) До начала работ по сбору массива лексических единиц, разработчику следует обратиться в депозитарный фонд с целью определения наличия зарегистрированных тезаурусов по заданной тематике.

При наличии таких тезаурусов проводят оценку возможности внедрения их в данной системе.

Первоначальный сбор лексики должен осуществляться выделением лексических единиц из представительной коллекции документов и/или запросов.

В полученный массив дополнительно включаются соответствующие тематике лексические единицы, выделенные в соответствии с методикой построения данного тезауруса из следующих источников:

Государственного рубрикатора НТИ;

рубрикатора соответствующей системы научно-технической информации;

энциклопедических, толковых и терминологических словарей и справочников;

терминологических стандартов;

классификаторов технико-экономической информации;

таблиц Универсальной десятичной классификации и других систем классификации;

нормативных (авторитетных) словарей коммуникативных форматов.

Наименования стран и других географических единиц должны соответствовать требованиям ГОСТ 7.67, наименования языков - требованиям ГОСТ 7.75.

3) Формирование словника ИПТ

В словник ИПТ могут быть включены следующие типы лексические единицы:

одиночные слова (существительные, прилагательные, глаголы, наречия);

именные словосочетания;

лексически значимые компоненты сложных слоев;

сокращения слов и словосочетаний.

Допускается включать словосочетания в словник, если в качестве опорного слова они содержат существительное и если выполнено одно из следующих условий:

- значение словосочетания не выводится из значений его компонентов.

- хотя бы один из компонентов словосочетания не употребляется в составе других сочетаний или употребляется всегда в другом смысле.

- для данного словосочетания в словнике существуют полные синонимы.

- данное словосочетание является устойчивым словосочетанием с именем собственным.

- отдельные слова словосочетания имеют слишком широкое значение.

- для данного словосочетания в словнике существует общепринятая аббревиатура.

- разбиение словосочетаний на отдельные компоненты приводит к потере важных для поиска семантических связей.

Словосочетания, которые не удовлетворяют перечисленным условиям, разбивают на компоненты.

**Дескриптор**

**– лексическая единица (слово, словосочетание) информационно-поискового языка, служащая для описания основного смыслового содержания документа или формулировки запроса при поиске документа (информации) в информационно-поисковой системе.**

**Аскриптор**

**– лексическая единица информационно-поискового тезауруса, которая в поисковых образах документов и запросов подлежит замене на дескриптор при поиске или обработке информации.**

4) Падежные формы одиночных существительных следует заменять формой именительного падежа. Падежные формы опорных существительных в словосочетаниях также заменяют формой именительного падежа с одновременным изменением падежной формы согласованных определений.

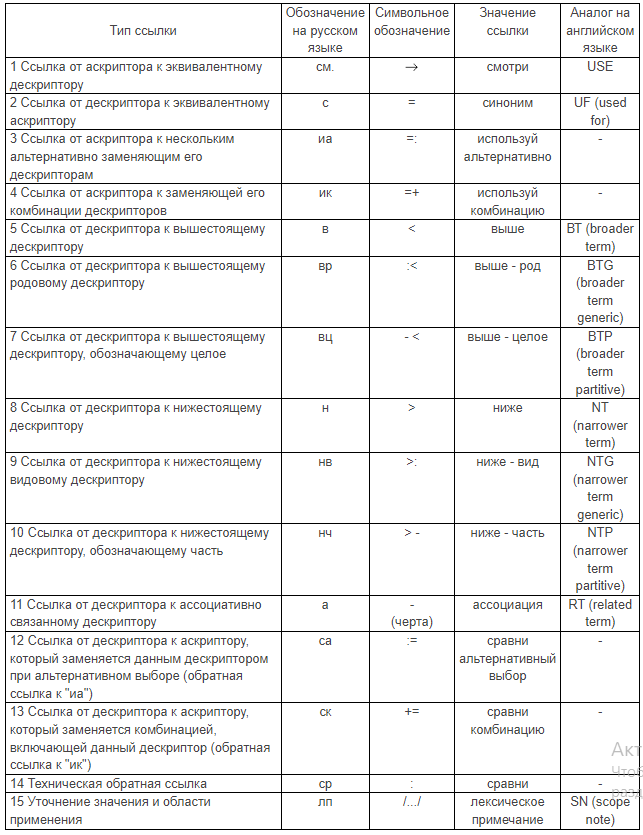
5) Формы прилагательных и причастий следует приводить к именительному падежу. Прилагательные и причастия в единственном числе приводят к форме мужского рода.

6) Глаголы рекомендуется заменять отглагольными существительными. Глаголы, включенные в словник, приводят к форме инфинитива.

7) Компоненты сложных слов, включаемые в словник ИПТ, должны быть широко употребляющимися частями сложных слов, общепринятых в данной специальности. Отсутствующий компонент сложного слова заменяют дефисом.

8) При наличии аббревиатур в словник должны быть включены также соответствующие им полные формы.

9) При построении словарных статей ИПТ лексическим единицам приписывают ссылки согласно таблице:



10) Ссылки отмечают связи данной ЛЕ с другими и являются результатом выполнения следующих операций:

- устранения неоднозначности лексической единицы;

- установления отношений эквивалентности;

- установления иерархических и ассоциативных отношений дескрипторов.

11) Неоднозначность лексических единиц устраняется релятором или лексическим примечанием в зависимости от характера неоднозначности.

12) Релятор является частью лексической единицы и поясняет ее значение, относя ее к определенной понятийной категории или предметно-тематической области. Допускается релятор представлять сокращением.

13) Лексические единицы объявляются эквивалентными в тезаурусе, если замена одной единицы на другую не приводит к изменению смысла текста, существенному для поиска информации.

14) Выбор дескрипторов проводится при формировании ИПТ первого типа.

Для выбора дескрипторов в классах эквивалентности выбирают одну лексическую единицу в качестве представителя всего класса.

Если в классе эквивалентности есть существительные, то представителем класса объявляется одно из них.

Если в классе эквивалентности более одного существительного, а также в тех случаях, когда существительные отсутствуют, критериями выбора представителя являются:

- полнота выражения значений данного класса лексических единиц;

- соответствие стандартам и рекомендациям по научно-технической терминологии;

- краткость и понятность;

- при прочих равных условиях наибольшая частота использования (распространенность) в документах и запросах.

15) Каждая лексическая единица получает статус дескриптора или аскриптора.

Статус аскриптора присваивают в следующих случаях:

- ЛЕ входит в класс эквивалентности, но не является предъявителем класса;

- значение ЛЕ целесообразно при поиске информации представить как пересечение значений других ЛЕ, получивших статус дескриптора;

- значение ЛЕ целесообразно при поиске информации представить как объединение значений других ЛЕ, несовместимых в одном поисковом образе.

В остальных случаях ЛЕ получает статус дескриптора.

16) Установление иерархических и ассоциативных отношений

Для дескрипторов устанавливаются парадигматические отношения, отражающие лексико-семантические связи между понятиями, выражаемыми дескрипторами. Связь указывают путем внесения в дескрипторную статью ссылки, включающей обозначение согласно таблице ссылок и связанный дескриптор.

Основными типами связей являются:

- род - вид,

- часть - целое,

- причина - следствие,

- сырье - продукт,

- административная иерархия,

- процесс - объект,

- функциональное сходство,

- процесс - субъект,

- свойство - носитель свойства,

- антонимия.

Отношения в тезаурусе могут быть paзделены на два класса: иерархические и ассоциативные.

Родовидовая связь устанавливается между двумя дескрипторами, если объем понятия нижестоящего дескриптора входит в объем понятия вышестоящего дескриптора.

Связь часть - целое устанавливается между двумя дескрипторами в том случае, если нижестоящий дескриптор обозначает компонент объекта, обозначаемого вышестоящим дескриптором.

Если для одного дескриптора можно указать более одного непосредственно вышестоящего дескриптора, то в иерархических отношениях должны быть установлены связи со всеми дескрипторами.

При установлении иерархических отношений должны быть указаны связи со всеми нижестоящими дескрипторами независимо от аспекта деления. Аспект деления может быть указан в примечании при ссылке.

При необходимости иерархические связи в дескрипторных статьях могут быть указаны на два и более уровней. В таких случаях уровни иерархии нумеруют относительно заглавной ЛЕ.

Ассоциативное отношение является объединением отношений, не входящих в иерархические отношения или в отношения синонимии. Допускается включать в ассоциативное отношение все виды отношений, кроме синонимии и отношения род - вид.

В целях обеспечения ведения ИПТ и индексирования документов ссылки, фиксирующие связи ЛЕ в словарных статьях, должных быть взаимными, т.е. для каждой ссылки, указывающей связь заглавной ЛЕ с другой ЛЕ, в словарной статье другой ЛЕ должна быть обратная ссылка. Если нецелесообразно использовать обратную ссылку при поиске информации, то следует применять технологическую обратную ссылку "сравни", обеспечивающую ведение ИПТ.

В дескрипторной статье указывают все синонимы заглавного дескриптора. Для каждого синонимичного аскриптора указывают один заменяющий дескриптор.

17) Построение лексико-семантического указателя

Лексико-семантический указатель является упорядоченной последовательностью словарных статей тезауруса и формируется путем расположения их в алфавитном порядке заглавных ЛЕ.

Устанавливается следующий приоритет символов при алфавитном расположении ЛЕ:

- пробел и знаки препинания, кроме дефиса (приоритеты всех знаков препинания и их сочетаний считаются одинаковыми),

- русские буквы,

- цифры,

- латинские буквы,

- буквы других алфавитов и специальные символы.

При этом ЛЕ, написанные через дефис, занимают то же алфавитное место, что и при слитном написании.

18) Построение систематического указателя

Систематический указатель является перечнем дескрипторов, сгруппированных согласно принятой рубрикации.

Систематические указатели подразделяют на три типа:

- тематические,

- категориальные,

- смешанные.

При построении систематического указателя тематического и смешанного типов в его тематической части следует использовать рубрики Межгосударственного рубрикатора НТИ или рубрикатора конкретной АСНТИ, совместимого с Межгосударственным рубрикатором НТИ.

При построении систематического указателя категориального и смешанного типов в его категориальной части следует использовать следующие общие категории:

- названия дисциплин и отраслей деятельности;

- предметы, материалы;

- методы, процессы, операции, явления;

- свойства, величины, параметры, характеристики;

- отношения, структуры, модели, законы, правила, абстрактные понятия.

В систематическом указателе категориального типа каждый дескриптор должен быть отнесен только к одной рубрике.

Внутри каждой рубрики дескрипторы располагают в алфавитном порядке.

19) Построение указателя иерархических отношений

Иерархический указатель отражает полную структуру иерархических отношений.

Иерархический указатель представляет собой перечень списков дескрипторов, причем каждый список начинается с дескриптора, не имеющего вышестоящих. После каждого дескриптора приведены непосредственно нижестоящие дескрипторы с указанием их уровня в иерархии путем применения нумерации либо графического обозначения уровня.

20) Построение пермутационного указателя

Пермутационный указатель предназначен для поиска ЛЕ по отдельным словам, входящим в текст ЛЕ, в том числе и по тем, которые стоят не в начале ЛЕ.

Пермутационный указатель представляет собой перечень списков, каждый из которых соответствует одному из знаменательных слов в составе ЛЕ и включает все ЛЕ, содержащие это знаменательное слово. Перечень списков упорядочен по алфавиту знаменательных слов, а внутри списки упорядочены по тексту ЛЕ.

Если число словосочетаний в ИПТ невелико, то допускается указывать пермутации словосочетаний непосредственно в лексико-семантическом указателе. При этом словосочетания в канонической форме приводятся в качестве дескрипторов, а другие варианты - в качестве синонимичных аскрипторов.

21) Структура, состав и форма представления тезауруса

В состав тезауруса входит вводная часть, основная часть (лексико-семантический указатель) и дополнительные части (систематический, пермутационный, иерархический и др. указатели и списки специальных категорий ЛЕ).

Обязательными составными частями являются вводная часть и лексико-семантический указатель.

Допускается в состав тезауруса вводить приложения, содержащие дополнительные сведения о разработке и использовании тезауруса.

Вводная часть включает титульный лист и введение.

На титульном листе должны быть приведены:

- наименование тезауруса, включающее термин "информационно-поисковый тезаурус" и указывающее область его применения;

- наименование организации-разработчика и дополнительные сведения об авторств;

- сведения о переиздании;

- место и год создания или издания.

Оформление титульного листа приведено в приложении А.

Введение должно содержать следующее:

- цель создания и область применения тезауруса, описанные кодами и наименованиями рубрик Межгосударственного рубрикатора НТИ;

- ссылки на источники, использованные для сбора лексики;

- ссылки на нормативные и методические документы, использованные при составлении;

- описание порядка составления тезауруса, включая обоснование представительности использованных источников лексики;

- описание состава и структуры;

- перечень отношений между ЛЕ и методические основания для их установления;

- перечень всех символов и специальных сокращений, допущенных для представления ЛЕ;

- порядок алфавитного расположения ЛЕ (расположение букв разных алфавитов, небуквенных символов и др.);

- количественные характеристики тезауруса (общее число статей; число дескрипторов, аскрипторов и др.);

- описание состава и формы представления дополнительных данных в словарных статьях;

- абзац следующего содержания: "Тезаурус подготовлен в соответствии с ГОСТ 7.25".

22) Форма представления дескрипторных и аскрипторных статей в лексико-семантическом указателе

Дескрипторная статья состоит из заглавного дескриптора, списка дескрипторов и аскрипторов, семантически связанных с ним, с обозначением видов связи.

Аскрипторная статья состоит из аскриптора и заменяющих его при обработке и поиске информации дескрипторов или комбинации дескрипторов.

23) Словарная статья может также включать:

- частоту использования дескриптора;

- кодовый номер дескриптора;

- код дескриптора по систематическому указателю;

- классификационные индексы;

- дополнительные семантические и лексикографические пометы;

- иноязычные эквиваленты.

Дополнительные данные располагают непосредственно за заглавным дескриптором и должны быть отделены от него.

Допускается указание дополнительных данных также при незаглавных ЛЕ.

24) В рамках дескрипторной статьи термины располагают в следующем порядке:

- заглавный дескриптор;

- дополнительные данные;

- лексическое примечание;

- аскрипторы или дескрипторы-синонимы;

- вышестоящие дескрипторы;

- нижестоящие дескрипторы;

- ассоциативные дескрипторы;

- дескрипторы, связанные другими видами отношений.

25) При печатном издании необходимо соблюдать следующие требования:

- дескрипторы должны быть выделены шрифтом или иным способом, позволяющим легко отличать их от аскрипторов и лексических примечаний;

- реляторы должны быть напечатаны в круглых скобках в конце дескриптора тем же шрифтом, что и остальной текст дескриптора, через пробел (допускается другой вид скобок, отличных oт косых);

- лексические примечания должны быть напечатаны после текста дескриптора в косых скобках через пробел или отдельной строкой, имеющей пометку "лп" (лексическое примечание);

- заглавная ЛЕ статьи должна выделяться левым сдвигом на расстояние, равное шести (или более) строчным буквам относительно начала незаглавной ЛЕ.