

ГОСТ 7.66-92 (ИСО 5963-85) СИБИД. Индексирование документов. Общие требования к координатному индексированию

ОКСТУ 0007

Дата введения 1993-01-01

Информационные данные

1. РАЗРАБОТАН И ВНЕСЕН Государственным комитетом СССР по науке и технологиям и Техническим комитетом ТК 191 "Научно техническая информация, библиотечное и издательское дело"

РАЗРАБОТЧИКИ

В.Н.Белоозеров, канд. филол. наук (руководитель темы); Н.Д.Кравченко, канд. пед. наук; И.В.Тростникова; Н.А.Сливницина; Г.Н.Хондкариан; В.Н.Казаков, канд. техн. наук

2. УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Постановлением Комитета стандартизации и метрологии СССР от 27.03.92 N 297

Настоящий стандарт разработан методом прямого применения стандарта ИСО 5963-85 "Документация. Методы анализа документов, определения их тематики и выбора терминов индексирования" с дополнительными требованиями, отражающими потребности народного хозяйства

3. Срок первой проверки - 1995 г.

Периодичность проверки - 5 лет

4. РАЗРАБОТАН ВПЕРВЫЕ

5. ССЫЛОЧНЫЕ НОРМАТИВНО-ТЕХНИЧЕСКИЕ ДОКУМЕНТЫ

Обозначение НТД, на который дана ссылка	Номер пункта, приложения
<u>ГОСТ 7.0-84</u>	Вводная часть
<u>ГОСТ 7.25-80</u>	4.2; 4.6
ГОСТ 7.26-80	Вводная часть
ГОСТ 7.27-80	Вводная часть; приложение 1
<u>ГОСТ 7.52-85</u>	Вводная часть; 5.7
<u>ГОСТ 7.59-90</u>	Вводная часть; приложение 1

Настоящий стандарт устанавливает общие требования к координатному индексированию документов, включая правила формирования поискового образа документа. Специфические требования к систематизации и предметизации документов - по ГОСТ 7.59. Форма представления поискового образа документа в коммуникативном формате МЕКОФ - по ГОСТ 7.52.

Стандарт распространяется на информационно-поисковые системы, в которых содержание документов представлено в сжатой форме лексическими единицами информационно-поискового языка. Стандарт не распространяется на формирование фактографических записей в фактографических базах данных.

Термины и определения - по ГОСТ 7.0, 7.26, 7.27, 7.59 и приложению 1.

Дополнительные требования, отражающие потребности народного хозяйства, приведены в приложении 1.

1. Общие положения

1.1. Процесс индексирования включает следующие этапы, которые осуществляют в указанной ниже последовательности:

анализ и определение содержания документа как объекта индексирования;

выбор понятий, характеризующих содержание документа;

выбор терминов индексирования для обозначения понятий;

формирование поискового образа документа из терминов индексирования.

Перечисленные этапы могут быть объединены в составе технологических процедур при условии надлежащего выполнения каждого из этапов.

1.2. Поисковый образ документа (ПОД) формируют из выбранных терминов

индексирования при помощи грамматических средств информационно-поискового языка (ИПЯ).

1.3. В процессе индексирования не рекомендуется описывать документ как физический объект (с точки зрения его формы, объема и пр.). Допускается отражать в ПОД подобную информацию, если она позволяет более точно установить соответствие документа информационной потребности пользователя системы.

2. Анализ документа

2.1. При анализе документа индексатору должна быть предоставлена возможность ознакомиться с документом в полном объеме. При невозможности исчерпывающего ознакомления с документом индексатор должен изучить имеющиеся текстовые части документа (основные источники индексирования):

справочный аппарат документа - заглавие (наименование), аннотацию, реферат, содержание (оглавление), предисловие, заключение и др.;

введение;

заголовки частей и глав;

первые фразы глав и параграфов;

иллюстрации, схемы, таблицы и подписи к ним;

слова и группы слов, которые в тексте подчеркнуты или выделены полиграфическими средствами.

Индексирование только по заглавиям является неполноценным. При индексировании по рефератам и аннотациям следует следить за адекватностью передачи в них содержания документа.

2.2. При анализе нетекстовых (аудиовизуальных и других) документов, которые помимо чтения требуют просмотра, прослушивания, испытания объекта в действии и других подобных процедур, допускается индексирование их по имеющемуся текстовому компоненту (наименованию, краткому описанию и т. п.), но и в этом случае индексатору должна быть предоставлена возможность полного ознакомления с документом, если текстовый материал представляется недостаточным.

3. Выбор понятий, характеризующих содержание документа

3.1. Число характеристик и понятий, отраженных в ПОД, определяет его полноту и является важнейшим показателем качества индексирования.

3.1.1. В ПОД необходимо отразить все понятия, которые могут иметь ценность для пользователей системы.

В документе может быть выявлено более одной темы из сферы интересов

пользователей. Эти темы должны рассматриваться отдельно.

3.1.2. Тематика, отражаемая при индексировании, не должна ограничиваться узкими рамками непосредственных интересов пользователей ИПС. Следует включать в ПОД также понятия, связанные с побочными аспектами документа (например, социальные и экономические аспекты научно-технических исследований).

3.1.3. При выборе понятий основным критерием является потенциальная ценность понятия для выражения содержания документа или для его поиска. При этом необходимо ориентироваться на типичные запросы к ИПС:

отбирать понятия, наиболее употребительные в коллективе пользователей ИПС;

уточнять состав лексики и грамматические правила ИПЯ на основе обратной связи с пользователями.

Изменения, вносимые в ИПЯ, не должны нарушать общую структуру и логику, заложенные при его создании.

3.1.4. Число терминов индексирования, приписываемых одному документу, определяется количеством сведений, содержащихся в документе. Ограничение числа терминов должно быть основано на содержательном отборе наиболее важных понятий.

3.2. Полнота индексирования, принятая в каждой ИПС, определяется ее функциональным назначением. Объем документа также сильно влияет на полноту индексирования. Необходимо учитывать указанные факторы и на их основе производить экспертный отбор понятий из документа, не стремясь включить в ПОД все упомянутые в нем понятия.

3.3. Специфичность ПОД определяется тем, в какой мере понятия документа нашли точное отражение терминами индексирования, и также является одним из параметров качества индексирования. Замена понятия термином, отражающим более широкое понятие, приводит к потере специфичности. Более широкие термины допускается использовать в особых случаях:

если излишне специфичный термин непонятен пользователям, особенно когда соответствующее понятие применяется только в пограничных областях деятельности;

если в документе понятие раскрыто недостаточно полно или является вспомогательным для изложения содержания документа.

3.4. Рекомендуется в каждой ИПС разрабатывать списки характеристик, которые признаются важными для отражения в ПОД. Для всех систем может быть рекомендован список указателей роли по ГОСТ 7.52. В зависимости от потребности конкретной ИПС этот список может быть как расширен, так и сокращен.

4. Выбор терминов индексирования

4.1. В процессе выбора терминов индексирования понятия, характеризующие содержание документа, представляют:

предпочтительными лексическими единицами (дескрипторами или ключевыми словами), выбранными по правилам конкретного ИПЯ;

терминами, отражающими новые понятия, проверив их точность и приемлемость по словарям, энциклопедиям, справочникам, классификационным таблицам, информационно-поисковым тезаурусам, терминологическим стандартам и другим источникам, признанным авторитетными в данной области.

4.2. Выбор терминов индексирования осуществляют на основе зарегистрированного (ГОСТ 7.25) или опубликованного информационно-поискового тезауруса, который используют при составлении запросов к ИПС.

При использовании тезауруса допускается сокращать число терминов, включаемых в ПОД за счет исключения общих понятий, которые могут быть привлечены на этапе поиска документа или на этапе составления поискового предписания на основании ссылок в статьях тезауруса.

4.3. Понятия, не представленные в словаре индексирования, но необходимые для формирования ПОД, выражают одним из двух способов:

новым специфическим термином, который включают в ПОД и в словарь;

более общим термином, имеющимся в ИПЯ; при этом специфический термин направляют в службу ведения ИПЯ в качестве кандидата на включение в словарь.

Новые понятия представляют наиболее близкими из существующих в ИПЯ лексических единиц, а также оценивают полезность включения новых терминов в словарь с точки зрения поиска.

4.4. При индексировании свободными ключевыми словами, взятыми из текста документа, они должны быть приведены к канонической форме по ГОСТ 7.25. Длину словосочетаний рекомендуется ограничивать двумя-тремя словоформами.

Схема индексирования с использованием информационно-поискового тезауруса приведена в приложении 2.

5. Формирование поискового образа документа

5.1. ПОД состоит из выбранных терминов индексирования, организованных с помощью грамматических средств ИПЯ данной ИПС.

5.2. В состав ПОД могут быть включены следующие категории данных, предусмотренные технологией индексирования конкретной ИПС:

степень нормализации терминов индексирования и применяемый для этого словарь;

индивидуальные характеристики термина индексирования;

связь терминов индексирования в синтаксических конструкциях ПОД.

Для включения в ПОД фактографических данных применяют грамматические

категории, указанные в разд. 6.

5.3. По степени нормализации различают два типа терминов координатного индексирования: дескрипторы и ключевые слова.

5.4. Термины индексирования должны быть представлены в ПОД в соответствии с орфографическими правилами используемого в системе естественного языка.

5.4.1. Дескрипторы допускается представлять условными кодами, которые указаны в используемом словаре индексирования. В этом случае ИПС должна обеспечивать автоматический поиск орфографических форм дескрипторов по их кодам.

5.4.2. Ключевые слова в многоязычных информационных системах, с ПОД на основе различных национальных языков, должны быть снабжены пометами о принадлежности к тому или иному естественному языку.

5.5. Индивидуальные характеристики терминов индексирования являются факультативными элементами ПОД и их используют для уточнения содержания документа, организации процедур информационного поиска или дальнейшей аналитико-синтетической обработки документов в системе.

К индивидуальным характеристикам относят данные о семантической и морфологической категории термина индексирования, его роли и информационном весе, способе получения и предполагаемом использовании.

5.5.1. Семантическая характеристика термина индексирования заключается в отнесении его к следующим лексикографическим категориям:

- 1) термин, выражающий научно-техническое понятие;
- 2) имя собственное, идентификатор;
- 3) наименование параметра;
- 4) значение параметра (выраженное текстом или именованной величиной);
- 5) числовое выражение;
- 6) обозначение единицы величины.

5.5.2. Морфологическая характеристика термина индексирования заключается в отнесении его к лексикографическим категориям:

- 1) производное слово;
- 2) сложное слово;
- 3) словосочетание;
- 4) аббревиатура;
- 5) фрагмент слова.

Морфологические характеристики используют в ПОД для реализации в ИПС

смыслового анализа лексических единиц на основе их формальных признаков.

5.5.3. Роль термина индексирования указывают в ПОД для уточнения места соответствующего понятия в содержании документа. Для этого особыми указателями роли, принятыми в ИПС, отмечают термины индексирования, отражающие следующие аспекты документа:

- 1) объект исследования, описания;
- 2) характеристики, свойства, параметры объекта;
- 3) методы и средства исследования, технологическую оснастку;
- 4) составные части, узлы, детали объекта;
- 5) область применения объекта (отрасль хозяйства, техники, науки);
- 6) назначение объекта;
- 7) цель исследования, разработки, описания;
- 8) результаты исследования, разработки.

5.5.4. Информационный вес термина индексирования отражает в ПОД важность данного понятия для данного документа. Число градаций информационного веса определяется потребностями конкретной ИПС. Следует различать:

- 1) понятия, выражающие главную тему документа;
- 2) понятия, выражающие побочные темы документа;
- 3) понятия, использованные в документе как вспомогательные для изложения его содержания.

Допускается использовать указатель отрицательного веса, которым помечают термины индексирования для указания на то, что данное понятие не рассматривается в документе.

5.5.5. Пометы, необходимые для указания на способ получения термина индексирования, используют для организации технологического процесса индексирования. Следует различать следующие пометы:

- 1) термин назначен по усмотрению индексатора, но отсутствует в документе;
- 2) термин введен в ПОД на основании связей, указанных в тезаурусе, но отсутствует в документе;
- 3) термин получен при автоматическом индексировании.

5.5.6. Пометы о предполагаемом использовании термина индексирования вводят в ПОД с целью выделить лексические единицы, подлежащие специальной обработке в процессах дальнейшей аналитико-синтетической переработки информации. Следует различать следующие пометы:

- 1) термин используется как предметная рубрика указателей:

2) при данном термине индексирования имеются фактографические данные, указанные в ПОД;

3) термин используется только как уточняющий определитель к другим терминам.

5.6. Термины индексирования в ПОД могут быть снабжены указателями связи, объединяющими их в синтаксические конструкции, которые отражают:

1) порядок следования и взаимное расположение терминов индексирования в документе;

2) смысловые связи понятий в документе;

3) парадигматические связи дескрипторов в тезаурусе.

Синтаксические конструкции рассматривают как цельные единицы ПОД наряду с терминами индексирования. Они могут быть объединены с другими синтаксическими конструкциями или с отдельными терминами индексирования в конструкции более высокого порядка.

Число уровней иерархии синтаксических конструкций определяется потребностями конкретных ИПС. Не следует применять конструкции четвертого и более высоких порядков.

Синтаксические конструкции могут быть охарактеризованы указателями веса, роли и предполагаемого использования аналогично индивидуальным терминам индексирования (см. пп.5.5.3, 5.5.4, 5.5.6).

5.7. Запись ПОД в памяти ИПС обусловлена принятым в ней способом кодирования с учетом требований настоящего раздела и ГОСТ 7.52.

6. Фактографическое индексирование документа

6.1. Фактографическое индексирование документа (ФИД) заключается в выявлении в документе и включении в ПОД данных, выражающих конкретные сведения (сообщения), имеющиеся в документе.

На основании результатов ФИД в фактографических ИПС формируются массивы сведений, в которых единицей информации является фактографическая запись.

6.2. ФИД предполагает формальное различие в ПОД двух категорий терминов индексирования, выражающих:

1) темы или объекты сообщения;

2) приписанные этим объектам свойства, являющиеся смыслом сообщения.

Соответствующие термины индексирования должны быть связаны друг с другом в синтаксическую конструкцию, объединяющую наименование объекта, его характеристики, их значения, единицы величины и отражающую смысловые связи понятий в документе.

Дополнительно такая синтаксическая конструкция может быть

охарактеризована:

- 1) показателем модальности;
- 2) условием истинности.

6.3. Показатель модальности фактографического сообщения определяет различие между сообщениями следующих типов:

- 1) наблюдаемый факт;
- 2) допускаемое значение;
- 3) требование стандарта;
- 4) плановый показатель;
- 5) запрет;
- 6) рекомендация;
- 7) предположение;
- 8) условие.

Если в информационной системе не используют показатели модальности, то все фактографические сообщения рассматривают как принадлежащие одной модальности, которая должна быть указана в эксплуатационной документации системы.

6.4. Условием истинности фактографического сообщения является другое фактографическое сообщение, связанное с первым в синтаксическую конструкцию вышестоящего уровня.

Например:

X = вес продукта

Z = 150 г.

V = влажность не более 45%,

где X - характеристика объекта,

Z - значение характеристики,

Y - условие истинности.

Фактографическое сообщение, являющееся условием истинности, должно иметь показатель модальности условия "если", например:

(вес продукта = 150 г) (если (влажность не более 45%)).

6.5. Термины индексирования, выражающие тему (объект) сообщения, относятся к категориям 1 или 2, указанным в п.5.5.1. При использовании категории 1 термину индексирования может быть дополнительно приписан показатель единичности или общности объекта (квантор).

Квантор общности используют в сообщениях, где выражено утверждение обо всех объектах, попадающих в объем соответствующего понятия.

Квантор единичности используют в сообщениях, где выражена информация о том объекте, входящем в состав данного понятия, который рассматривается в данном документе.

6.6. Термины индексирования, выражающие свойства объектов, которые составляют смысл сообщения, могут быть выражены лексическими единицами категорий 1, 2, 3 (см. п.5.5.1) или параметрической конструкцией (см. п.5.6).

6.7. Параметрическая конструкция должна состоять из двух формально выраженных частей: наименования параметра и перечня значений параметра (см. п.6.8), которые объединены в одну синтаксическую конструкцию.

6.8. Перечень значений в параметрической конструкции должен включать набор значений параметров и указание об альтернативности или одновременности (симультантности) значений.

Набор значений задают перечислением или указанием двух предельных значений, между которыми располагаются значения, принимаемые параметром (интервалом значений). При задании интервала значений формально указывают, которое из значений является начальным и конечным для интервала значений, а также входят ли граничные значения в указанный интервал. Одно из граничных значений интервала может отсутствовать, если значение параметра ограничено только с одной стороны.

Указание об одновременности используют, когда у одного объекта сообщения наблюдаются все заданные значения параметра. Указание об альтернативности используют, когда параметры одного объекта сообщения должны быть выбраны из числа заданных.

6.9. Значения параметра могут быть представлены синтаксической конструкцией из двух терминов индексирования - числового выражения и наименования единицы величины - при необходимости производить операции расчета или численного сравнения.

7. Автоматизированное индексирование

7.1. Целью автоматизации индексирования является минимизация материальных и человеческих ресурсов, затрачиваемых на процедуру индексирования, а также достижение стабильности и единообразия ее результатов.

7.2. Автоматизированное индексирование (АИ) осуществляют по:

- 1) тексту первичного документа.
- 2) заглавию и аннотации или реферату документа;

АИ по тексту первичного документа должно включать процедуру сжатия ПОД.

7.3. С использованием вычислительной техники осуществляют следующие содержательные этапы АИ:

- 1) выявление информативных частей документа;
- 2) идентификация слов текста и приведение их к нормализованному виду (морфологический анализ и синтез);
- 3) формирование списка ключевых слов исходного текста;
- 4) подбор дескрипторов по тезаурусу;
- 5) формирование ПОД.

7.4. Выявление информативных частей документа

Технология AI должна предусматривать идентификацию и предоставление индексатору или программе индексирования наиболее информативных фрагментов документа из списка указанных в п.2.1. Могут быть предусмотрены алгоритмы выявления информативных фрагментов по другим формальным критериям, а также по решению специалиста-индексатора.

7.5. Идентификация слов текста

7.5.1. Процесс идентификации слов текста должен включать: отождествление словоформ одного слова и определение информативных слов текста.

При этом может быть необходимо использование интеллектуальных процедур для решения таких задач, как выявление и обработка синтаксических конструкций, выявление и разрешение омонимии.

7.5.2. Для идентификации слов текста используют машинные словари (словари основ, парадигм, словосочетаний и т.д.). Словари должны быть представлены в базе данных системы и обеспечены средствами визуализации и ведения.

7.6. Формирование списка ключевых слов текста

7.6.1. В процессе формирования списка ключевых слов текста проводится синтаксический анализ текста с учетом правил сочетаемости грамматических категорий данного естественного языка.

7.6.2. Синтаксический анализ текста решает задачи:

- 1) разделение текста на фрагменты по заданным критериям;
- 2) установление синтаксических зависимостей между словоформами текста;
- 3) отождествление словосочетаний;
- 4) нормализация выявленных ключевых слов.

7.7. Автоматическое формирование ПОД

7.7.1. В процедуре AI допускается формирование ПОД из свободных ключевых слов или дескрипторов информационно-поискового тезауруса, используемого в данной области.

7.7.2. При AI дескрипторами информационно-поискового тезауруса на этапе формирования ПОД происходит замена ключевых слов на дескрипторы,

указанные в тезаурусе.

7.7.3. При формировании ПОД из дескрипторов возможно обогащение ПОД за счет пополнения вышестоящими терминами информационно-поискового тезауруса.

7.7.4. Процедура АИ должна предусматривать включение в ПОД типовых грамматических средств (см. разд. 5).

7.7.5. К системам АИ предъявляются следующие требования:

1) модульность построения, т.е. такая внутренняя организация лингвистического и программного обеспечения системы, при которой процедуры решения отдельных задач АИ реализуются с помощью самостоятельных блоков или модулей;

2) ориентация на типовые программные и технические средства;

3) соответствие действующей нормативно-методической документации по координатному индексированию.

Приложение 1 (справочное). Термины и определения

1. Автоматизированное индексирование - индексирование, технология которого предусматривает использование формальных процедур, осуществляемых с помощью вычислительной техники, и может включать применение интеллектуальных процедур при принятии основных решений о составе поискового образа.

2. Автоматическое индексирование - составление поискового образа с использованием только формальных процедур обработки текста документа или запроса, осуществляемых средствами вычислительной техники.

3. Информативное слово - слово или словосочетание в тексте документа или запроса, которое несет в нем существенную смысловую нагрузку.

4. Контролируемое индексирование - индексирование, при котором предусмотрена замена информативных слов текста дескрипторами, указанными в определенном информационно-поисковом тезаурусе или другом словаре индексирования.

5. Координатное индексирование - индексирование, цель которого состоит во всестороннем отражении содержания документа или запроса путем включения в поисковый образ всех необходимых для этого терминов индексирования.

6. Лексическая единица (ЛЕ) ИПЯ - последовательность символов, слово, словосочетание, фрагмент слова или условное обозначение, которая рассматривается в данном ИПЯ как элементарная единица, используемая для представления в поисковых образах документов или запросов определенного понятия, объекта или значения параметра.

7. Свободное индексирование - индексирование, технология которого не предусматривает замену информативных слов текста в соответствии с рекомендациями специального словаря индексирования.

8. Специфический термин - информативное слово, в наибольшей степени отражающее содержание документа, использование которого отличает данный документ от других тематически близких документов.
9. Специфичность индексирования - характеристика качества индексирования, определяемая отношением числа специфических терминов и фактографических сведений к числу неспецифических терминов в поисковом образе.
10. Полнота индексирования - степень отражения в поисковом образе содержания документа и (или) запроса, определяемая как отношение числа специфических терминов и фактографических сведений, включенных в поисковый образ, к числу таковых терминов и сведений, имеющих в тексте документа или запроса.
11. Фактографическое индексирование - индексирование, предусматривающее отражение в поисковом образе документа конкретных сведений (сообщений), являющихся смыслом данного документа.

Приложение 2 (справочное). Схема индексирования по информационно-поисковому тезаурусу

1. Изучить документ и составить перечень существенных для его содержания понятий с учетом специфики ИПС.
2. Рассмотреть первое понятие
3. Найти в тезаурусе лексическую единицу, отражающую данное понятие. Если таковой нет, перейти к п.11.
4. Если найденная лексическая единица - аскриптор, заменить ее указанным в ссылке дескриптором (или комбинацией дескрипторов).
5. Рассмотреть ссылки, указанные в тезаурусе для данного дескриптора (дескрипторов).
6. Проверить, не являются ли указанные в ссылках дескрипторы более специфичными для выражения данного понятия. Если да, то перейти к п.10.
7. Записать найденные лексические единицы в поисковый образ, снабдив их необходимыми грамматическими показателями по правилам данного ИПЯ.
8. Проверить, имеются ли еще не отраженные в поисковом образе понятия из документа и рассмотреть следующее понятие. Перейти к п.3.
9. Если список понятий документа исчерпан, окончить работу.
10. Заменить исходный дескриптор более специфичными согласно указанию ссылки в тезаурусе. Перейти к п.7.
11. Найти в тезаурусе дескрипторы, совместное включение которых в поисковый образ отражает данное понятие. Если таковых нет, перейти к п.12, если есть - перейти к п.5.

12. Установить термин, выражающий понятие и удовлетворяющий требованиям к дескрипторам по ГОСТ 7.25.
13. Направить найденный термин в службу ведения ИПЯ в качестве кандидата на включение в тезаурус. Перейти к выполнению п.7.
14. Конец.

Блок-схема индексирования по информационно поисковому тезаурусу показана на чертеже.

Блок-схема алгоритма индексирования

