МЕТОДИЧЕСКИЕ АСПЕКТЫ СОКРАЩЕНИЯ ОБЪЕМА ТЕКСТОВЫХ ДОКУМЕНТОВ НА ПРЕДПРИЯТИЯХ МИКРОЭЛЕКТРОННОЙ ПРОМЫШЛЕННОСТИ

Черников Б.В., Борисова Е.А.,
ООО «Газпром ВНИИГАЗ», г. Москва
Российский экономический университет им. Г.В. Плеханова, г. Москва
bor-cher@yandex.ru, lb20062006@yandex.ru
Кремер Е.А.
Московский институт электронной техники, г. Москва
kremerea@gmail.com

Аннотация. Увеличение количества текстовых документов, применяемых на предприятиях, влечет за собой рост трудозатрат, направленных на создание этих документов. Применение лексикологического синтеза может позволить исправить сложившуюся ситуацию. В работе рассмотрена методика сокращения объема текстовых документов на предприятиях микроэлектронной промышленности.

Ключевые слова: лексикологический синтез; методика; хранение; слабоформализуемый документ; индексная последовательность.

Введение

В настоящее время документооборот на предприятиях в большинстве случаев осуществляется в бумажном виде, несмотря на широкое распространение персональных компьютеров и применение

на предприятиях систем электронного документооборота. Данный факт обуславливает необходимость наличия архива бумажных документов значительных объемов.

Преобладание бумажных документов над электронными копиями определяется рядом причин, к которым можно отнести следующие:

– требования документов, регламентирующих документооборот внутри предприятия (помимо внутренних стандартов предприятия сюда входят государственные и ведомственные стандарты);
– недостаточное распространение электронных средств защиты текстовых документов, таких как электронная подпись;
– требования предоставлениям документов регулирующим органам, вышестоящим организациям и партнерам в бумажном виде;
– высокая юридическая значимость.

Процесс формирования документов является обязательным элементом функционирования предприятия, что накладывает на него определенные условия.

К процессу документирования информации предъявляются требования, к которым следует отнести следующие:

• максимальная формализация;
• минимальные временные затраты на создание документа;
• максимально компактный объем формируемого файла;
• автоматизация формирования конкретного документа при слабой его формализации.

Данные требования соответствуют российским и европейским стандартам [1, 2]. Одним из таких является международный стандарт MoReq 2010 [2], который накладывает дополнительные условия на весь процесс формирования документов. Внедрение данного стандарта требует повышение удобства процесса создания документов.

Документы, используемые для информационного обмена на предприятиях микроэлектронной промышленности, в подавляющем большинстве являются слабоформализуемыми, поскольку их содержание сильно зависит от конкретной ситуации, однако при этом необходимо обеспечить возможность учета всех нюансов документируемой ситуации.

Слабоформализуемые документы – полнотекстовые, табличные либо смешанные документы, содержание которых существенным образом связано с произвольной, меняющейся от конкретной ситуации структурой. Это документы, обладающие достаточно высокой степенью вариативности. В связи с этим содержательная структуризация слабоформализуемых документов может требовать детализации как взаимосвязи, так и взаимной зависимости композиции текста вплоть до атомарных значений – фрагментов фраз, слов, и даже частей отдельных слов [3].

Актуальность исследований в области документооборота, документирования, хранения и передачи информации обуславливается приоритетной программой правительства «Информационное общество» [4], а также общим увеличением количества и объема хранимых электронных документов.

Целью данного исследования является разработка методики сокращения размера слабоформализуемых документов при хранении на предприятиях микроэлектронной промышленности при помощи лексикологического синтеза.

1 Лексикологический синтез

Одним из эффективных путей для исполнения требований, приведенных выше, является автоматизированное формирование документов, в котором применяется технология лексикологического синтеза [3].

Лексикологический синтез – формирование текстовых фрагментов путем создания фраз на основе набора опорных слов, который формируется по результатам глубокого анализа текста документа путем связывания текстовых фрагментов с конкретным опорным словом, входящих в состав фраз или выражений формируемого текста.

В результате унификации содержания документов формируется набор формулировок, которые могут присутствовать в документе. Каждой из них ставится в соответствие одно слово, которое является основным и однозначно определяет наличие конкретной фразы в тексте документа. Данные слова называются опорными и являются основой лексикологической схемы (лексонтологии) конкретного документа. Полный перечень опорных слов комплекса документов с учетом вида внедряемой в документ информации образует лексикологическое дерево.

Формирование документа происходит на основе внедрения необходимых формулировок путем выбора опорных слов, соответствующих данным формулировкам. В процессе создания документа в результате выбора того или иного опорного слова формируется индексная последовательность. Данная индексная последовательность и является, по сути, сформированным документом. Для прочтения данного документа необходимо, чтобы на принимающей стороне было установлено необходимое программное обеспечение, связанное с таким же лексикологическим деревом, которое было использовано при создании документа. Размер данной индексной последовательности при таком способе формирования документа оказывается во много раз меньше, чем непосредственно документ, созданный традиционными способами. И в то же время такая последовательность содержит в себе все характерные особенности содержания сформированного документа. Кроме того, такой способ создания документов позволяет снизить вероятность появления ошибок, а также сократить время и трудозатраты, необходимых для создания текстового документа. При этом дополнительно обеспечивается простота чтения готового документа, как если бы это был документ, созданный традиционным методом набора текста. Это является важными факторами на предприятиях микроэлектронной промышленности.

Основой лексикологического синтеза является тот факт, что каждая область и сфера работы на предприятии сопровождается конкретным комплексом документов. Любой документ, описывающий ситуацию в исследуемой отрасли, содержит переменную и постоянную информацию. При анализе текстового документа можно выделить постоянную информацию, характерную именно для данного вида документов. К постоянной информации добавляется переменная информация, которая может принадлежать конечному множеству вариантов, если текст заранее унифицирован. Поскольку множество вариантов конечно, то, объединив их, можно информацию отнести к разряду переменной унифицированной. При создании документа постоянная информация вносится автоматически, а переменная унифицированная информация внедряется после выбора нужных вариантов из сохраненного множества.

Лексикологический синтез – формирование текстовых фрагментов путем создания фраз на основе набора опорных слов, который формируется по результатам глубокого анализа текста документа путем связывания текстовых фрагментов с конкретным опорным словом, входящих в состав фраз или выражений формируемого текста.

В результате унификации содержания документов формируется набор формулировок, которые могут присутствовать в документе. Каждой из них ставится в соответствие одно слово, которое является основным и однозначно определяет наличие конкретной фразы в тексте документа. Данные слова называются опорными и являются основой лексикологической схемы конкретного документа. Полный перечень опорных слов комплекса документов образует лексикологическое дерево.

Формирование документа происходит на основе внедрения необходимых формулировок путем выбора опорных слов, соответствующих данным формулировкам. В процессе создания документа в результате выбора того или иного опорного слова формируется индексная последовательность. Данная индексная последовательность и является, по сути, сформированным документом. Для прочтения данного документа необходимо, чтобы на принимающей стороне было установлено необходимое программное обеспечение, связанное с таким же лексикологическим деревом, которое было использовано при создании документа. Размер данной индексной последовательности при таком способе формирования документа оказывается во много раз меньше, чем документ, созданный традиционными способами. И в то же время такая последовательность содержит в себе все характерные особенности содержания сформированного документа.

2 Методика сокращения объема документов

Рассмотрим методику сокращения объема документов с помощью лексикологического синтеза (рис. 1).

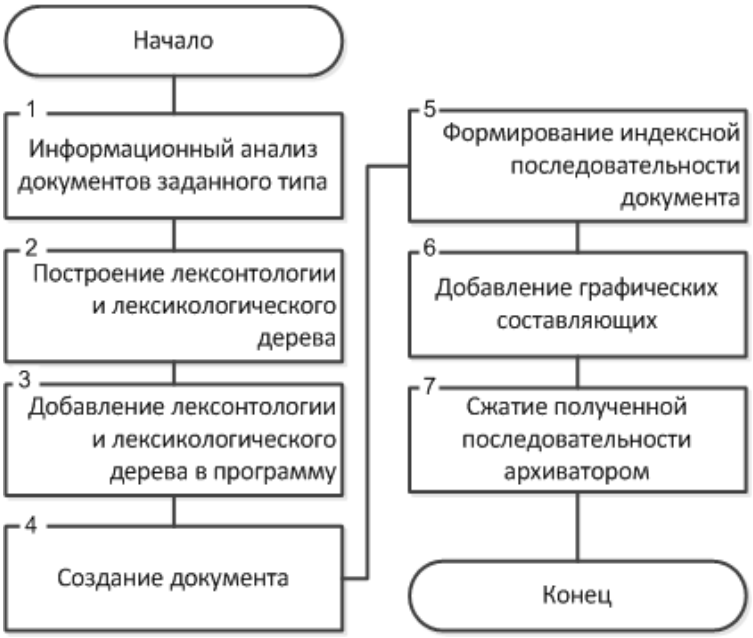

Рис. 1. Методика сокращения объема документов

1. На первом этапе необходимо провести глубокий информационный анализ того вида документов [5], создание которых планируется автоматизировать. Количество и наличие информации в документе изменяются. На присутствие или отсутствие информации в документе влияют различные факторы. Ситуаций, в которых создаваемые документы будут совпадать, практически не бывает.

Для возможности применения лексикологического синтеза необходимо проанализировать структуру информации, которая содержится в документе. Вся информация традиционно делится на постоянную и переменную. Постоянная информация – неизменная информация, которая используется в течение длительного периода времени без каких-либо изменений. Переменная информация отражает фактические количественные и качественные характеристики деятельности предприятия, которые необходимо закрепить в документе. К переменной информации можно отнести всю информацию, которую необходимо вводить в документ при каждом его заполнении.

Информация, содержащаяся в документе, подвергается более глубокой классификации с разделением на четыре категории [6]:

- унифицированная постоянная информация;
- унифицированная переменная информация;
- переменная вводимая информация;
- неунифицированная информация.

Унифицированная постоянная информация подготавливается заранее, хранится в базе данных и автоматически внедряется в документ при его создании. Данный тип информации включает в себя постоянную и редко меняющуюся информацию. Унифицированная переменная информация содержит стандартизированные и формализованные данные, хранится в базе данных, и предоставляется путем выбора требуемых формулировок. К данному типу информации относятся формулировки, предлагаемые создателю документа для выбора при формировании конкретных частей документа. Переменная вводимая информация содержит конкретизирующие данные, свойственные данному экземпляру документа. Неунифицированная информация включает произвольные формулировки, вводится с клавиатуры.

Такое разделение информации по сравнению с классическим делением только на постоянную и переменную позволяет уменьшить объем неунифицированной информации, а также детально характеризовать каждую конкретную ситуацию на предприятии.

2. На втором этапе строится лексикологическая схема (лексонтология) и лексикологическое дерево документа заданного вида.

На основе лексического анализа выделяются фрагменты документа, которые определяются конкретным словом либо словосочетанием, обозначающим наличие конкретной формулировки в документе. Такие слова называются опорными. Основным критерием включением какого-либо

слова или словосочетания во множество опорных слов является его однозначное понимание и применение. В качестве критериев выделения опорных слов выделяют следующие [3]:

- фонетический – опорное слово должно соответствовать звуковому строю языка документа;
- фоносемантический – звучание опорного слова должно вызывать ассоциации, непосредственно связанные с формируемым документом определенного вида;
- морфологический – составные, гибридные или сокращенные опорные слова, аббревиатуры или их идентификаторы должны как в полном объеме, так и составными его частями, вызывать ассоциации, связанные с создаваемым документом;
- лексический и семантический – выбираемые или назначаемые опорные слова, их совокупность или идентификатор должны базироваться на лексике документа;
- лексикографический – опорные слова, их совокупность или идентификатор должны легко читаться, смысл их должен быть понятен для любого составителя данного документа.

Количество опорных слов должно быть необходимым и достаточным. Если опорных слов будет мало, то это, безусловно, уменьшит время формирования документа, однако при этом придется укрупнять фрагменты текста, соответствующие тому или иному опорному слову, что может негативно сказаться на вариативности заполнения документа. Малый объем множества опорных слов оправдан в случае устойчивых слов и словосочетаний, которые могут определяться всего лишь одним словом (это можно наблюдать, например, в организационно-распорядительных документах). Увеличение мощности множества опорных слов позволит увеличить вариативность описываемых в документе ситуаций, но при этом увеличит время создания документа.

Поэтому выделяют четыре критерия создания совокупности опорных слов [3]:

- лексико-смысловое единство;
- сбалансированная насыщенность множества опорных слов;
- смысловая ценность, обеспечивающая легкое запоминание смысловых цепочек опорных слов;
- функционально-стилевая принадлежность опорных слов к предметной области создаваемых документов, которая отражает специфику терминов, применяемых на предприятии.

Далее сформированное множество опорных слов применяется для создания лексикологической схемы (лексонтологии) и дерева документов. Лексонтология документа представляет собой модель взаимной связи опорных слов, входящих в состав сформированного множества и используемых в процессе создания документов данного вида с учетом вариативности отдельных экземпляров [3]. Лексонтология позволяет отобразить связь между опорными словами при последовательном создании документа, являясь своеобразной инфологической или онтологической моделью комплекса документов. Для учета типа внедряемых фрагментов, соответствующих лексонтологии, используется структура типа дерево, она позволяет отобразить взаимные связи слов с учетом различных возможных вариантов внедрения фрагментов в документ. Совместный учет категорий классификации информации и взаимосвязей опорных слов позволяет сформировать лексикологическое дерево. Пример лексикологического дерева приведен на рис. 2.

3. На третьем этапе построенная лексонтология и лексикологическое дерево импортируются в исполняемую программу, с которой будут работать составители и создатели документов.

4. На четвертом этапе происходит создание документа.

5. На пятом этапе формируется индексная последовательность на основе внесенной в документ информации в соответствии с алгоритмом, приведенным на рис. 3.

Фиксация индексной последовательности, соответствующей выбираемым опорным словам, осуществляется пошагово в рамках организуемого цикла выбора опорных слов. В случае отсутствия в лексонтологии унифицированного варианта формулировки, определяемого опорным словом, в индексную последовательность внедряется вводимый неунифицированный фрагмент. По завершении процесса формирования документа в разделе подписей фиксируется индекс подписи должностного лица (исполнителя документа), который также внедряется в индексный информационный пакет.

Рис. 2. Пример лексикологического дерева приказа об увольнении [3]

Рис. 3. Алгоритм формирования индексной последовательности [6]

Принципиальным отличием данного варианта от классического способа формирования индексной последовательности является использование двоичных кодов, где «0» – опорное слово не выбрано, а «1» – опорное слово выбрано. В рамках каждого уровня фиксируется своя двоичная последовательность, которая переводится в десятичную систему, получая тем самым одно десятичное число на уровень. Все уровни располагаются друг за другом через пробел. В случае наличия неунифицированных фрагментов они располагаются после десятичного числа соответствующего уровня и располагаются внутри символов «(» и «)», а после завершения формирования последовательности конкатенируются в одну строку, после чего сжимаются с помощью выбранного алгоритма архивации, а потом добавляются в конец индексной последовательности, а все предыдущие неунифицированные фрагменты удаляются, и заменяются номерами, соответствующими номерам в конкатенированной строке.

Рассмотрим условный пример фиксации фрагмента индексной последовательности при формировании технологической инструкции с помощью классического способа формирования индексной последовательности и с помощью измененного способа формирования индексной последовательности, который приведен в [3]. Лексикологическая схема технологической инструкции представлена на рис. 4.

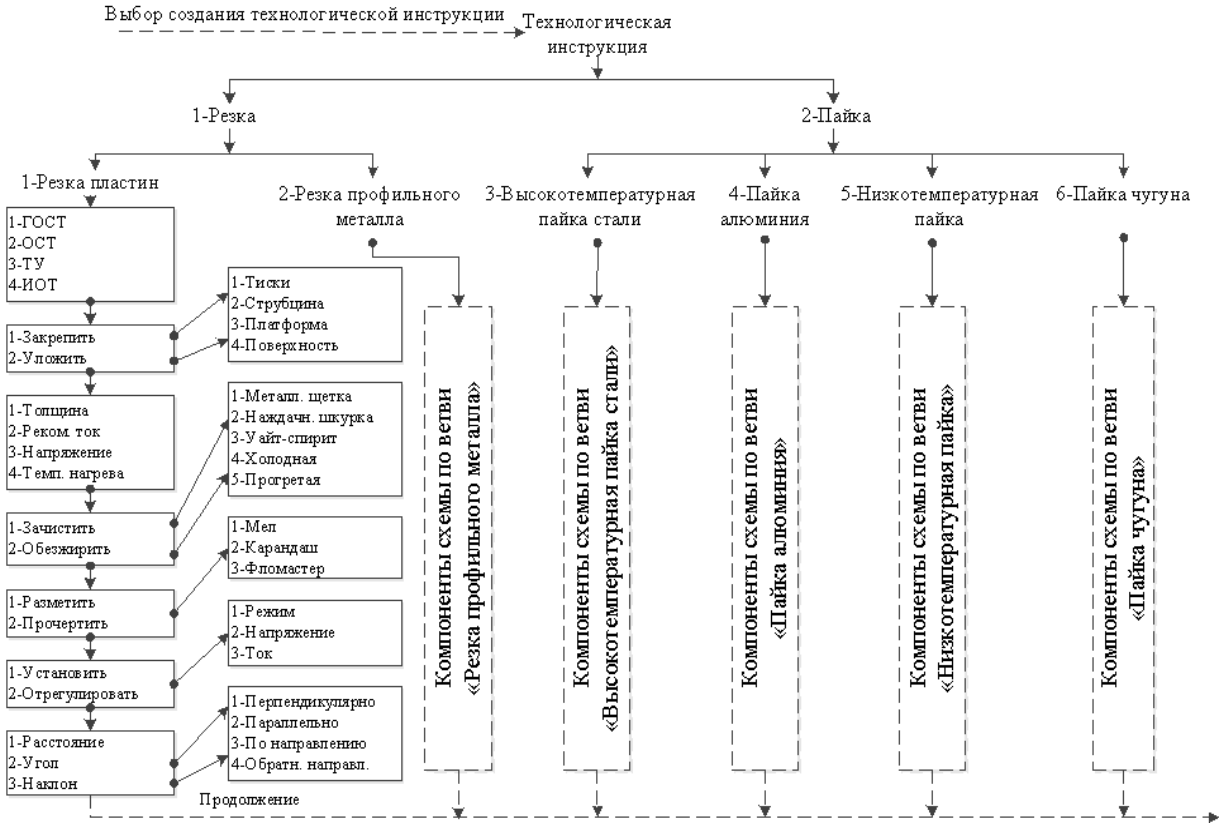

Рис. 3. Лексонтология технологической инструкции

Таблица 1. Сравнение классического и нового способа формирования последовательности

№ уровня	Классический способ	Новый способ
1	1-3+«Плазма-12м-02»	4(«Плазма-12м-02»)
2	2-1	1
3	3-1	1
4	4-1+4	9
5	5-1:1+2	1(3)
6	6-1:1-2;2-4;4-6+2:3-5;5-7;6-8+3:140-260;140-170;150-170	7(1-2 2-4 4-6;3-5 5-7 6-8;140-260 140-170 150-170)
7	7-1:1+5	1(17)
8	8-1:1	1(1)
9	9-1:1+2	1(3)
10	A-1:3-5+2-1+3:5-10+4	7(3-5;1;5-10;4)

В итоге получим следующую последовательность для классического способа:

1-3+«Плазма-12м-02» 2-1 3-1 4-1+4 5-1:1+2 6-1:1-2;2-4;4-6+2:3-5;5-7;6-8+3:140-260;140-170;150-170 7-1:1+5 8-1:1 9-1:1+2 A-1:3-5+2-1+3:5-10+4.

Для нового способа последовательность выглядит следующим образом:

4(«Плазма-12м-02») 1 1 9 1(3) 7(1-2 2-4 4-6;3-5 5-7 6-8;140-260 140-170 150-170) 1(17) 1(1) 1(3) 7(3-5;1;5-10;4)

После этого архивируем строку «Плазма-12м-02», так как она является единственной неунифицированной информацией и добавляем полученный результат в конец индексной последовательности. А вместо «Плазма-12м-02» записываем «1».

Очевидно, что новый способ дает выигрыш в длине последовательности, а соответственно и объеме хранимого документа.

6. На шестом этапе к сформированной индексной последовательности добавляются графические составляющие, такие как логотипы, картинки и т.п. В конец последовательности дописывается индекс графической составляющей.

7. На седьмом этапе полученная последовательность дополнительно сжимается архиватором.

Выводы

В результате исследования получены следующие результаты:

1. Разработана методика сокращения текстовых документов и доказана ее эффективность на примере технологической инструкции.
2. Разработанная методика определяет четкую последовательность действий для сокращения объема текстового документа и позволяет применять ее на предприятиях микроэлектронной промышленности.
3. В результате исследования были выявлены следующие проблемы, на которые будет направлено дальнейшее исследование:
   - возможен большой объем неунифицированной информации в индексной последовательности, что существенно увеличит ее размер – следует проанализировать способы сокращения ее объемов;
   - необходимость возможности подстановки своих неиндексированных графических материалов в последовательности, что влечет необходимость совершенствования метода добавления индекса графических составляющих.

Литература

1. Model Requirements for the Management of Electronic Records (MoReq 2). – France: European Communities. – 2008. – 212 p.
2. Modular Requirements for Records Systems (MoReq 2010). – CECA-CEE-CEEA, Bruxelles-Luxembourg. – 2011. – P. 520.
3. *Черников Б. В.* Лексикологический синтез документов в комплексах информационных систем. – М.: ИД «ФОРУМ». – 2017. – 336 с.
4. Государственная программа Российской Федерации «Информационное общество (2011-2020 годы)» (утв. Постановлением правительства Российской Федерации от 15.04.2014 № 313) // Собрание законодательства РФ. – 05.05.2014, № 18 (2 ч.), ст. 2159.
5. *Черников Б. В., Кремер Е. А.* Анализ информации текстовых документов предприятий микроэлектронной промышленности// Современные наукоемкие технологии. – 2018. – № 5, с. 168-172.
6. *Черников Б. В., Кремер Е. А.* Сокращение объема индексной последовательности при лексикологическом синтезе слабоформализуемых документов// Информатизация и связь. – 2018. – № 6, с. 58-64.
7. ГОСТ 3.1105-2011. Единая система технологической документации (ЕСТД). Формы и правила оформления документов общего назначения. – М.: Стандартинформ, 2012. – 24 с.