ТЕХНОЛОГИЯ АВТОМАТИЗИРОВАННОГО ФОРМИРОВАНИЯ СЛАБОФОРМАЛИЗУЕМЫХ ДОКУМЕНТОВ

Карминский А.М.
МГТУ им. Н.Э.Баумана, г. Москва
karminsky@mail.ru
Черников Б.В.
ООО «АНТ-Информ», г.Москва
bor-cher@yandex.ru

Ключевые слова: слабоформализуемые документы, лексикологический синтез, опорные (ключевые) слова

Введение

Процесс выработки управленческих решений включает в себя несколько этапов, один из которых заключается в сборе и оценке полученных сведений. Эта информация должна быть представлена в соответствии с установленными требованиями. Необходимо также, чтобы собранные сведения были в максимальной степени удобны для восприятия на следующих этапах. Поэтому возникает проблема способа закрепления информации в документе, то есть процесса ее документирования. Процесс формирования документов различного характера является обязательным элементом при возникновении необходимости сохранения информации о каком-либо процессе или событии.

Процесс документирования информации должен отвечать ряду требований, к которым должны быть отнесены следующие:

- данные должны быть максимально формализованы в целях обеспечения автоматизированной обработки сведений, содержащихся в документе;
- создание документов должно занимать минимум времени при сохранении требований к информации, необходимых для принятия управленческого решения.

Дополнительные требования к процессу документирования устанавливает принятый в начале 2008 г. стандарт MoReq2 [1], рекомендации которого, несомненно, будут приниматься во внимание при разработке систем электронного документооборота на российском рынке. Одним из требований, выдвигаемых MoReq2, является удобство процесса создания документов.

Актуальность современных исследований в области совершенствования средств и методов формирования документов возрастает в связи с процессом реализации в государстве Федеральной целевой программы «Электронная Россия» и началом формирования электронного правительства, которое должно быть создано как на региональном, так и не федеральном уровне. Важность значения данного направления подтверждается тем, что вопросы его развития рассматриваются на самом высоком уровне, включая федеральный Совет по развитию информационного общества и Государственный совет.

Управленческие документы в подавляющем большинстве можно охарактеризовать как слабоформализуемые, поскольку при высокой степени вариативности содержания, зависящего от конкретной ситуации, они должны в целом обеспечивать возможность фиксации всех возможных нюансов, соответствующих сфере применения.

Слабоформализуемые документы – это полнотекстовые, табличные либо смешанные документы, содержание которых существенным образом связано с произвольной, меняющейся от конкретной ситуации структурой. Эти документы обладают достаточно высокой степенью вариативности. Поэтому 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.

1. Современное состояние процесса создания документов

Традиционный способ формирования текстовых документов предполагает ввод информации с клавиатуры путем прямого набора текста. В этом случае трудозатраты на формирование документов оказываются весьма существенными. Кроме того, набору текста часто сопутствует появление

орфографических и синтаксических ошибок, вызванных, например, техническими погрешностями или недостаточно высоким уровнем грамотности исполнителя документа. Дополнительной проблемой часто становятся возможные погрешности в оформлении документов, которые автор, создавая тот или иной документ, вынужден изучать, чтобы соблюсти установленные нормативы, определяющие форму и содержание документа.

В настоящее время выделяются, в основном, два вида унификации текстов: типизация и трафаретизация. В случае типизационного подхода создаются сборники типовых текстов, на основе которых формируется текст конкретного документа. Анализ показывает, что в документах существуют элементы представления информации, которые могут быть представлены типовыми текстами. Раньше этот метод (метод стандартных фраз) применялся в организационно-распорядительной документации для стандартизации деловой переписки.

При трафаретизации текстов информацию подразделяют на постоянную и переменную. Постоянная информация вносится в бланк документа при его изготовлении, а переменная – в процессе составления конкретного документа.

Одним из направлений, позволяющих сократить время формирования документов и снизить вероятность появления ошибок, является использование специализированных карт, содержание которых основано на применении формализованной информации. Однако подобный способ предполагает использование ограниченного диапазона элементов, входящих в состав документа. В этой связи он находит достаточно широкое применение в основном при формировании учетных документов. Другим направлением является использование шаблонов документов в распространенных текстовых редакторах. К сожалению, этот способ так же не свободен от недостатков, в качестве одного из которых следует привести невозможность вариации размеров отведенных текстовых полей в широких пределах, что актуально при существенных изменениях объемов фрагментов документа.

2. Сущность лексикологического синтеза документов

Естественным оказывается стремление пользователей, с одной стороны, снизить вероятность появления ошибок и сократить трудозатраты, необходимые для формирования текстовых документов, а с другой стороны, – сохранить легкость чтения сформированного документа на уровне составленного традиционным способом прямого ввода информации при соблюдении требований по оформлению документа.

Весьма эффективным путем удовлетворения этих запросов является применение способа автоматизированного формирования документов, основанного на принципе лексикологического синтеза [2]. Информация, содержащаяся в подготавливаемом тексте документа, подвергается более глубокой классификации, что позволяет при создании текста документа сократить объем неунифицированной информации, формировать текст документа, наиболее точно характеризующий каждую конкретную управленческую ситуацию.

Критерий эффективности должен характеризовать формирование текста любого документа из множества используемых в обеспечении процесса управления в соответствующих областях деятельности. Создаваемый документ должен обладать необходимой полнотой содержащейся в нем информации и формироваться в приемлемые сроки. При этом необходимо учитывать вариативность формирования текстов в зависимости от конкретной управленческой ситуации. Следовательно, основным критерием, используемым при решении исследуемого комплекса задач, является полнота содержащейся в нем информации при допустимом времени формирования документа с учетом вида создаваемого документа. Принимая во внимание, что каждый документ содержит совокупность смысловых и содержательных компонент-разделов, математическое представление основного критерия может иметь следующий вид:

$$\boldsymbol{P} = \sum_{m=1}^{M} p_m \rightarrow \max \left| \forall R_{ij} \; ; T_{\text{форм}} \leq T_{\text{доп}} \right. , \quad (1)$$

где

$\boldsymbol{P}$ – критерий формирования содержания документа;

p_m – полнота представления m-го компонента документа;

R_{ij} – вариантность i-го вида документа для j-й управленческой ситуации;

$T_{\text{форм}}$ – время формирования документа;

$T_{\text{доп}}$ – допустимое время формирования документа.

Несмотря на то, что основной критерий характеризует качество формируемых документов, целесообразно рассмотреть и измеряемый критерий, который может иметь принципиальное значение при оценке эффективности исследуемого способа формирования документов. В качестве подобного критерия целесообразно выбрать время формирования документа

$$T_{\text{форм}} \rightarrow \min , \quad (2)$$

минимизация которого позволит снизить расходы на его создание и, следовательно, себестоимость документа.

Основой лексикологического синтеза является тот факт, что определенная сфера человеческой деятельности сопровождается унифицированным набором документов. Любой документ, фиксирующий информацию об определенной управленческой ситуации, содержит информацию двух классов – постоянную и переменную. Из текста документа можно выделить характерную для него постоянную информацию. Постоянная информация дополняется переменной, причем в заранее унифицированных текстах документов переменная информация может принадлежать предопределенному множеству вариантов. Если это множество вариантов предварительно, путем экспертной оценки, собрать воедино, переменная информация может быть отнесена к разряду переменной унифицированной. Современные инструментальные средства позволяют осуществить хранение этих, заранее зафиксированных, вариантов формулировок. В последующем при создании документа реализуется автоматическое внедрение в документ постоянной информации, а также автоматизированное включение переменной информации, выбранной исполнителем документа из сохраненного множества.

Под термином «лексикологический синтез» следует понимать в данном случае формирование текстовых фрагментов компьютерной системой путем синтеза (создания) фраз на основе использования набора опорных (ключевых) слов, комплектуемого по результатам предварительно проведенной содержательной унификации документа, с автоматическим связыванием фрагментов и отдельных слов текста в соответствии с правилами орфографии и лексикологии. Использование понятия «лексика» в данном случае является оправданным, поскольку оно определяет «словарный состав языка, набор используемых в данном языке ключевых слов» [3]. Особенностью применения этого термина является его распространение лишь на содержание конкретного вида документа, поэтому речь следует вести только о формировании необходимого набора слов, определяющего возможность составления определенного документа.

В ходе унификации документов по содержанию формируется набор (по возможности – полный, то есть предусматривающий все возможные варианты) формулировок, которые могут присутствовать в документе. При этом необходимо учитывать разнообразие информации, которая может содержаться в каждом отдельном экземпляре формируемого документа. Следовательно, по отношению к отдельно взятому экземпляру документа сформированный набор формулировок может и должен быть избыточным, то есть содержать даже большее количество фрагментов текста, чем это необходимо при составлении единичного экземпляра документа. Сформированный набор формулировок сохраняется в упорядоченном состоянии в виде индексированной совокупности элементов (например, в виде базы данных). Каждой формулировке ставится в соответствие основное слово, выбор которого однозначно определяет наличие конкретной формулировки в документе. Такие слова называются опорными и образуют основу лексикологической схемы формируемого документа. Сопоставление индексов сохраненного набора формулировок опорным словам позволит обеспечить однозначную взаимосвязь этих элементов. Полный перечень опорных слов образует лексикологическое дерево документа, «прохождение» по ветвям которого обеспечит выбор формулировок, используемых в документе. При этом выбор тех или иных опорных слов будет означать необходимость внедрения в документ конкретных вариантов текстовых фрагментов.

Фактически, текст документа формируется путем выбора необходимых заготовок из числа сохраненных формулировок. Задача программных средств заключается в обеспечении необходимой связи между словами в используемых фразах, что может потребовать некоторого изменения отдельных слов в формулировках в целях их согласованного (с точки зрения правил синтаксиса) применения. К сожалению, в отдельных случаях определение полного набора формулировок

товых фрагментов может оказаться затруднительным. В этой связи целесообразно усматривать возможность ввода свободных формулировок прямым набором текста. Имеющийся применения лексикологического синтеза слабоформализуемых документов показывает, что при точно внимательном исследовании возможной структуры документа объем подобных ментов, как правило, оказывается незначительным и не оказывает существенного влияния на водительность труда сотрудника, выполняющего задачу формирования документа. Структура есса подготовки документа к автоматизированному формированию приведена на рис. 1

ис. 1. Структура процесса подготовки формирования слабоформализуемого документа

оцесс инициируется принятием решения о необходимости автоматизированного рования слабоформализуемого документа с помощью лексикологического синтеза.

соответствии с задачами, решаемыми в конкретной сфере деятельности, устанавливается ность реквизитов, расположенных в определенной последовательности, и определяется ложение каждого элемента информации. Определяются зоны унифицированной формы нта, предназначенные для закрепления ее в технических средствах хранения документов, а для нанесения специальных изображений (например, логотипа организации). Устанавливается реквизитов, необходимых и достаточных как для идентификации автора официального нта, так и для придания документу юридической силы [4]: наименование автора документа, умента, дата, номер, место составления или издания. В дальнейшем такие реквизиты в целях ии времени постоянно вносятся в бланк документа программными средствами. После ния унификации формы документа производится ее сохранение в базе знаний, содержащей рованные формы документов, подлежащих автоматизированному формированию.

Процесс инициируется принятием решения о необходимости автоматизированного формирования слабоформализуемого документа с помощью лексикологического синтеза.

Затем проводится классификация информации слабоформализуемого документа. В отличие от традиционной системы классификации информации (на постоянную и переменную) для лексикологического синтеза необходим более глубокий методологический подход, обеспечивающий классификацию информации в информационные потоки следующих типов:

- унифицированная постоянная информация, внедряемая в документ автоматически с использованием базы знаний. К этому типу относится постоянная (например, наименование документа) и редко меняющаяся (наименования организации и структурного подразделения, список персонала и т.п.) информация;
- унифицированная переменная информация, содержащая стандартизированные и формализованные данные, хранимая в базе знаний и вводимая при формировании документа путем выбора требуемых формулировок;
- переменная вводимая информация, представляющая конкретизирующие сведения, как правило, для конкретного экземпляра документа (например, табличные данные, отдельные фамилии и т.п.) и вводимая с клавиатуры непосредственно при подготовке документа;
- неунифицированная информация, содержащая свободные формулировки и вводимая прямым набором с клавиатуры.

Необходимость выделения в отдельную категорию унифицированной переменной информации вызвана тем, что формулировки (слова, словосочетания, фразы), входящие в состав информационного потока этого типа, являются платформой, на основе которой синтезируется совокупность опорных слов. Именно эти формулировки, сопоставляемые с опорными словами, должны сохраняться в базе данных для автоматического внедрения в документ при выборе соответствующего опорного слова. Особенность переменной вводимой информации заключается в наличии условного форматирования у данного типа данных, что должно предусматривать программную проверку вводимых сведений.

Следующей операцией является синтез совокупности опорных слов. Опорное слово выбирается на основе метода анализа языковых конструкций, свойственных данному управленческому документу. В перечне возможных формулировок определяются слова (или их совокупность), привязка к которым позволяет однозначно выбрать требуемую формулировку Каждое опорное слово в зависимости от содержания конкретного экземпляра документа может соответствовать как целой фразе, так и определенному словосочетанию или даже отдельному слову в документе. Так, например, выбор опорного слова «норма» в медицинском протоколе может означать необходимость внедрения в документ целой фразы, характеризующей соответствие описываемого параметра принятым нормам. Однако в большинстве случаев опорные слова соответствуют более коротким формулировкам, присутствие которых в документе необходимо в соответствии с описываемой ситуацией. В связи с этим количество опорных слов в значительной степени зависит от вариативности документа и требуемой степени гибкости лексикологического дерева, которая определяется детальностью анализа при проведении синтеза совокупности опорных слов.

Взаимная зависимость опорных слов в совокупности определяет последовательность обхода маршрута формирования документа. На основе анализа структуры документа выявляются основные разделы, которые должны или могут присутствовать в документе. Условные наименования таких разделов составляют основу синтезируемой совокупности опорных слов. В рамках каждого зафиксированного раздела документа выявляются составные элементы, которые должны или могут входить в состав раздела (слово, фраза, текстовый фрагмент). Для каждого подобного составного элемента определяется опорное слово (или их совокупность), выбор которого в последующем однозначно будет определять внедрение в документ соответствующей компоненты. Если фрагмент текста документа содержит значительное количество строк и всегда присутствует в документе в строго определенной последовательности построения предложений, то данный фрагмент текста определяется одним опорным словом. Однако в случаях, когда текст документа формируется из предложений, не фиксированных в строго определенной последовательности, и в каждом заново создаваемом документе наблюдаются вариации построения текста, опорных слов будет столько, сколько необходимо для однозначного определения каждого конкретного предложения или словосочетания.

Синтез лексикологического дерева, проводимый далее, предназначен для определения маршрута хода опорных слов на основе учета их взаимосвязей. Полный перечень опорных слов с их имосвязями образует лексикологическое дерево, «прохождение» по ветвям которого обеспечивает бор всех необходимых формулировок, используемых в документе. При этом выбор тех или иных рных слов означает необходимость внедрения в документ конкретных вариантов текстовых гментов. Структура лексикологического дерева сходна с композицией текста документа.

Степень ветвления лексикологического дерева зависит от объема множества вариаций текста кумента, определяемых его сложностью и различием документируемых ситуаций. В качестве рного слова могут выступать различные части речи, определяющие сущность предписываемого йствия. Генерация лексикологического дерева осуществляется при соблюдении критериев равления лексическими конструкциями. Опорное слово должно быть уникальным для конкретной струкции, а при необходимости уточняться другими опорными словами, иначе выбор требуемого стового фрагмента может быть определен неверно. Уточнение одного опорного слова другим разует их иерархическую подчиненность в структуре лексикологического дерева. При этом зывание опорных слов имеет значение только в смысле порядка их следования в маршруте обхода ксикологического дерева при формировании документа. Проведение цикла выбора определенной следовательности опорных слов означает формирование экземпляра документа конкретного вида с. 2, где утолщенной линией условно указан маршрут выбора опорных слов φ, отсекающего ряд ковых ветвей).

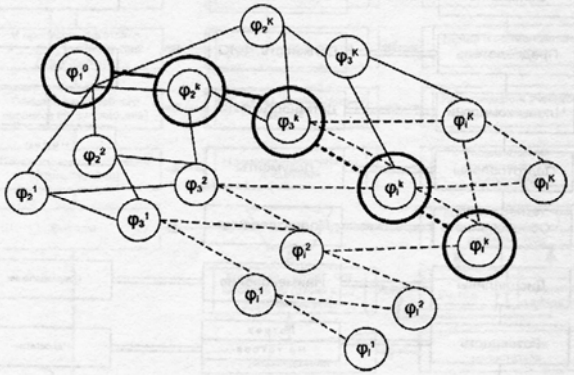


Рис. 2. Модель формирования документа при использовании дерева с отсечением

Референтная взаимосвязь опорных слов при необходимой мощности их множества позволяет рмировать модель создания документа D, принадлежащего к множеству документов еделенного вида $\boldsymbol{D}^{\mathrm{B}}$, при наличии существенных вариаций в рамках отдельных экземпляров:

$$\forall D \in \boldsymbol{D}^{\mathrm{B}}\ \exists x \in \Psi^{\mathrm{B}} = \{\varphi_1, \varphi_2, \ldots, \varphi_b, \ldots, \varphi_{\mathrm{B}}\}$$

$$F(x)=\Phi(\varphi_1) \vee \Phi(\varphi_2) \vee \ldots \vee \Phi(\varphi_b) \vee \ldots \vee \Phi(\varphi_{\mathrm{B}}), \quad (3)$$

$^{\mathrm{B}}$ – множество опорных слов φ, сформированное для документов данного вида;
– набор опорных слов, используемый при создании конкретного экземпляра документа данного ида;
(x) – комплекс фрагментов документа, каждый из которых связан с определенным опорным овом φ.

Пример лексикологического дерева для автоматизированного формирования акта готовности кафедры высшего учебного заведения к новому учебному году приведен на рис. 3. Формулировки сохраняются в базе знаний с индексированием, соответствующим опорным словам.

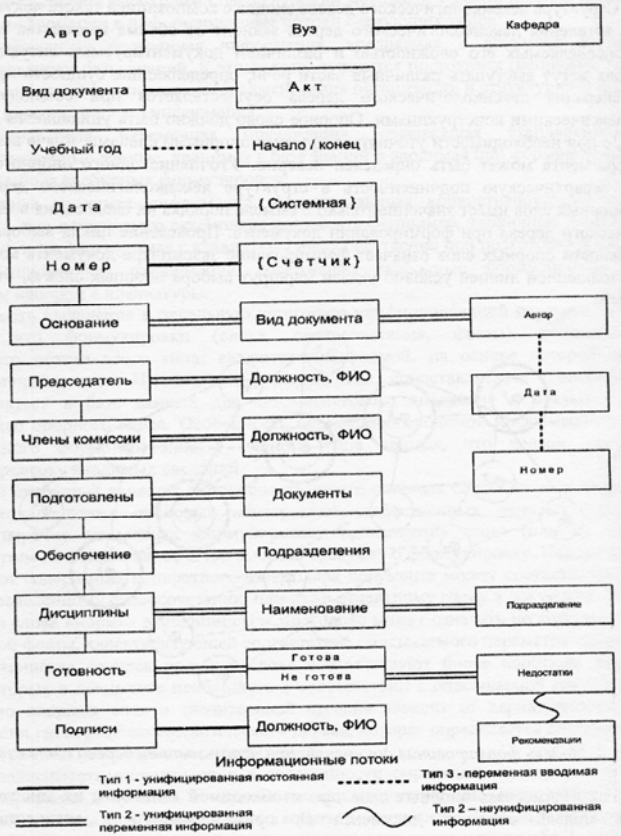


Рис. 3. Пример лексикологического дерева документа

На следующем этапе формируется информационный алгоритм документа, который позволяет учесть вид функциональных связей отдельных опорных слов и текстовых фрагментов (формулировок) документа, и определяет способ внедрения фрагмента в документ.

Установление типов связи опорных слов лексикологического дерева и текстовых фрагментов, однозначно определяемых данным опорным словом, осуществляется с помощью построения информационного алгоритма. Связь опорного слова с унифицированным текстовым фрагментом изображается графически с помощью заранее выбранных обозначений. Пример информационного алгоритма для документа «Акт готовности кафедры к новому учебному году», сформированный для автоматизированного формирования в высшем учебном заведении, приведен на рис. 4.

Содержанием следующего этапа является синтез оперативного алгоритма (процедуры матизированного формирования документов), структура которого представляет собой основу для дующего создания соответствующей компьютерной программы.

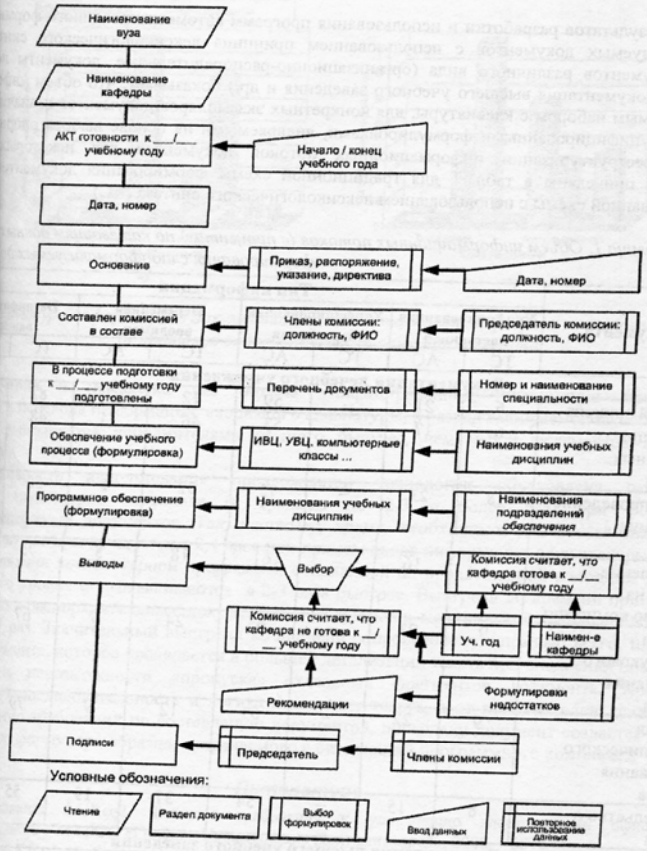


Рис. 4. Пример информационного алгоритма

процедуре учитываются взаимосвязи блоков и фрагментов будущей программы, определяется ательность работы пользователя при автоматизированном формировании документа. разрабатываемой процедуры является управление процессом последовательного выбора слов. Кроме этого, процедура организует пошаговое наращивание текста документа путем нения функционально связанных текстовых фрагментов (формулировок). Разработка ры должна быть ориентирована на применение персональной компьютерной техники. ра процедур может выполняться в нотациях и с помощью инструментов построения ных алгоритмов.

Завершающим этапом разработки технологии является создание компьютерной программы формирования документа, который впоследствии создается в интерактивном режиме.

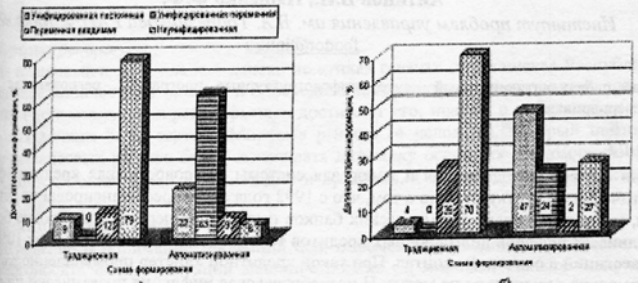
Заключение

Анализ результатов разработки и использования программ автоматизированного формирования слабоформализуемых документов с использованием принципа лексикологического синтеза при создании документов различного вида (организационно-распорядительные, документы лечебного учреждения, документация высшего учебного заведения и др.) показывает, что объем информации, вводимой прямым набором с клавиатуры, для конкретных экземпляров документов незначителен по сравнению с унифицированными формулировками, внедряемыми на основе выбора опорных слов. Результаты реструктуризации информационных потоков документов для некоторых систем документации приведены в табл. 1 для традиционной схемы формирования документов (ТС) и автоматизированной схемы с использованием лексикологического синтеза (АС).

Таблица 1 Объем информационных потоков (в процентах) по категориям данных и схемам формирования слабоформализуемых документов

Вид документа	Тип информации							
	Унифицированная постоянная		Унифицированная переменная		Переменная вводимая		Неунифицированная	
	ТС	АС	ТС	АС	ТС	АС	ТС	АС
Документация лечебного учреждения								
Выписной эпикриз	6	21	–	59	12	7	82	13
Документация консилиумных совещаний	6	17	–	53	46	14	48	16
Заявки на проведение обследования пациента	18	23	–	72	32	1	50	4
Представление пациента на врачебную комиссию	7	16	–	61	18	5	75	18
Протокол ультразвукового обследования пациента	9	23	–	65	24	7	67	5
Протокол эндоскопического обследования пациента	9	22	–	63	12	9	79	6
Свидетельство о болезни	8	15	–	54	37	19	55	12
Документация высшего учебного заведения								
Акт готовности кафедры к новому учебному году	4	47	–	24	26	2	70	27
Отчет о практиках	8	80	–	7	54	5	38	8
Отчет о работе ГЭК	9	61	–	15	46	15	45	9
Отчет о работе кафедры в учебном году	3	56	–	7	17	3	80	34
Протокол заседания ГЭК	18	32	–	46	27	–	55	22

Пример графического отображения результатов реструктуризации информационных потоков ...дены на рис. 5.



...5. Структура информации в документах: а) акт готовности кафедры к новому учебному году; б) протокол эндоскопического осмотра пациента

...еличение в несколько раз объема унифицированной информации предопределяет ...ность перевода информации, вводимой с клавиатуры, в разряд выбираемой на основе опорных ...что выполняется исполнителями намного быстрее, чем набор фрагментов документа с ...туры.

...ффективность использования предложенной технологии обусловлена существенным ...ием трудозатрат персонала (в среднем до 2-3 и более раз) при формировании ...формализуемых документов. Так, например, время, необходимое для составления типового ...го контракта, сокращается с 2,4 часа при прямом вводе информации с клавиатуры до 1,1 часа ...пользовании компьютерной программы, основанной на предложенной методике. Документы ...ального уровня подготавливаются в 2-3 раза быстрее. Выигрыш во времени при подготовке ...ционно-распорядительных документов и документов медицинской направленности может ...ть 5-7 раз. Значительный выигрыш во времени сопровождается, помимо всего, повышением ...а документа, которое проявляется в большей детализации текста при внедрении описательных ...нтов и невозможности «пропуска» отдельных фрагментов документа. Кроме того, ...ивается последовательность и логика изложения текста, а от исполнителей документов не ...ся знание требований по составлению документов, поскольку документ создается на основе ...енного формуляра-образца, сохраняемого в базе знаний программного комплекса.

Литература

Спецификация MoReq2. Типовые требования к управлению электронными официальными документами. – М.. РОО «Гильдия Управляющих Документацией», 2008. – 287с.

Черников Б.В. Способ автоматизированного лексикологического синтеза документов // Патент России № 2253893. 2005.

Першиков В. И., Савинков В. М. Толковый словарь по информатике. – М.. Финансы и статистика, 1991 – 543с.

Борисов А.Б. Большой экономический словарь. – М.. Книжный мир, 2003. – 895с.